Towards Dynamic Intelligent Support for Collaborative Problem Solving

Sidney D'Mello[1], Angela E. B. Stewart[1], Mary Jean Amon[1], Chen Sun[2], Nicholas Duran[2], & Valerie Shute[3]

[1] University of Colorado Boulder, Boulder CO 80309, USA
[2] Arizona State University, Glendale AZ 85306
[3] Florida State University, Tallahassee FL 32306
3
sidney.dmello@colorado.edu

Abstract. We discuss progress towards the design of collaborative interfaces that automatically assess key facets of collaborative problem solving (CPS) and intervene accordingly. Our work is grounded in a generalized theoretical model of CPS including three major facets (constructing shared knowledge, negotiation/coordination, and maintaining team function), subfacets, and verbal and nonverbal indicators. We report results of two studies that validated the model followed by speech and language processing techniques to automate the assessment of the CPS facets. We conclude by discussing future plans on how to incorporate the models in next-generation CPS interfaces that support dynamical assessment and intelligent intervention.

Keywords: Assessing Collaborative Problem Solving; Natural Language Processing; Machine Learning.

1 Introduction

It is widely acknowledged that collaborative problem solving (CPS) is an essential 21st century skill in our increasingly connected and globalized world [1]. Yet, we know precious little about how to define, measure, and help develop this skill, especially in the context of STEM learning. By increasing our basic understanding of effective CPS processes, we can take a step towards designing next-generation collaborative learning environments that aim to make CPS more enjoyable and effective. Accomplishing this vision requires: (1) identification of effective CPS processes (or facets); (2) automatically monitoring the core CPS processes to enable intervention; and (3) designing and testing the efficacy of intelligent collaborative interfaces with dynamic intervention and/or after-action feedback and reflection. Here, we describe progress on the first two of these components and sketch out ideas for the third component.

2 Collaborative Problem Solving Model and its Validation

We synthesized previous research on CPS to construct a generalized CPS competency model (i.e., skills and abilities) from existing frameworks, such as ATC21S [2] and PISA [3] along with some classic work on CPS [4, 5]. Since our model is intended to be generalizable, we validated it using data from two very different studies.

2.1 CPS Model

Our model consists of the following core facets: (1) constructing shared knowledge (expresses one's own ideas and attempts to understand others' ideas); (2) negotiation/coordination (achieves an agreed solution plan ready to execute); and (3) maintaining team function (sustains the team dynamics). Each facet has two sub-facets, which in turn, have multiple verbal and nonverbal indicators as shown in Table 1.

2.2 Model Validation

We validated our competency model in two studies [6]. In Study 1, 11 triads of middle school students (8th-9th graders) played Physics Playground (PP) face-to-face for three hours. This is a 2D educational video game that was developed to support and measure students' learning of *conceptual physics* [7]. It focuses on Newton's laws of force and motion, mass, gravity, potential and kinetic energy, and conservation of momentum. Problems (or levels) in PP require students to guide a green ball to a red balloon. The primary way students move the ball is by creating *agents*, simple machines of force and motion (i.e., ramps, levers, pendulums, and springboards), drawn with colored lines using the mouse, that "come to life" on the screen. For example, Figure 1 (ultimate pinball level) shows a sample problem where the student must draw a carefully constructed ramp (in purple) to lead the falling ball along a path to the balloon. Students receive silver trophies for any solution to a problem but earn gold trophies for elegant solutions involving a limited number of objects created and used to solve the problem (the threshold varies but is typically < 3).

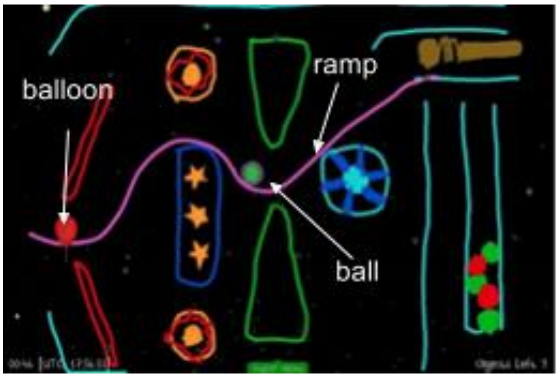


Fig. 1. A level in Physics Playground.

Table 1. Generalized competency model composed of facets, sub-facets, and indicators.

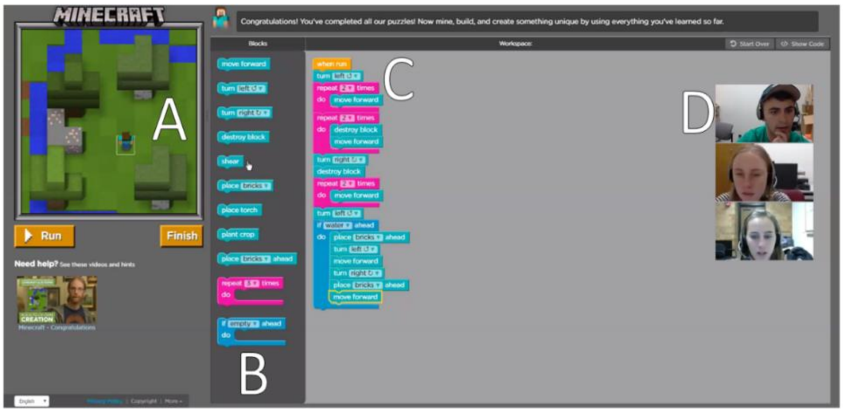
Facet (Sub-facet)	Indicators
Constructing shared knowledge: expresses ideas and attempts to understand others' ideas	
Shares understanding of problems and solutions	• Talks about specific topics/concepts and ideas on problem solving • Proposes specific solutions • Talks about givens and constraints of a specific task • Builds on others' ideas to improve solutions
Establishes common ground	• Recognizes and verifies understanding of others' ideas • Confirms understanding by asking questions/paraphrasing • Repairs misunderstandings • Interrupts or talks over others as intrusion (R)
Negotiation/Coordination: achieves an agreed solution plan ready to execute	
Responds to others' questions/ideas	• Does not respond when spoken to by others (R) • Makes fun of, criticizes, or is rude to others (R) • Provides reasons to support/refute a potential solution • Makes an attempt after discussion
Monitors execution	• Talks about results • Brings up giving up the challenge (R)
Maintaining team function: sustains the team dynamics	
Fulfills individual roles on the team	• Not visibly focused on tasks and assigned roles (R) • Initiates off-topic conversation (R) • Joins off-topic conversation (R)
Takes initiatives to advance collaboration processes	• Asks if others have suggestions • Asks to take action before anyone on the team asks for help • Compliments or encourages others

Note. "R" next to an indicator means that it is reverse coded.

Below is an excerpt of an exchange between two participants (Player A and Player C) during gameplay, along with tags for the relevant indicators.

- **Player C**: “What if you grabbed it upwards. And then drew a pendulum, knock it out. But you drew like farther out, the pendulum” (Proposes specific solutions)
- **Player A**: “I have an idea. Wait, which direction should I swing?” (Confirms understanding by asking questions/paraphrasing)
- **Player C**: “Swing from here to here.” (Proposes specific solutions)
- **Player A**: “Nope, then it would just fly to the spider.” (Provides reasons to support/refute a potential solution)

In Study 2, 37 undergraduate triads played Minecraft-themed Hour of Code for 20 minutes using videoconferencing. This is an online resource for students grades two and above to learn basic computer programming principles in an hour. It uses a visual programming language, Blockly (https://developers.google.com/blockly/), to interlock blocks of code (such as loops). Blockly eliminates syntax errors by only interlocking syntactically correct blocks, allowing students to focus on the coding logic and programming principles (see Figure 2).



Congratulations! You've completed all our puzzles! Now mine, build, and create something unique by using everything you've learned so far.

```
when run
turn left ↺
repeat 2 times
do  move forward
repeat 2 times
do  destroy block
    move forward
turn right ↻
destroy block
repeat 2 times
do  move forward
turn left ↺
if water ahead
do  place bricks ahead
    turn left ↺
    move forward
    turn right ↻
    place bricks ahead
    move forward
```

Fig. 2. Minecraft-themed Hour of Code. Students can watch their code run (A), choose from a code bank of possible blocks (B), generate code (C) and see their teammates (D).

Below is an excerpt of an actual exchange between all three participants (Players A, B, and C) during gameplay, along with the relevant indicators.

- **Player C:** “Yeah I think so. Cuz we’ll fall in, right?” (Provides reasons to support/refute a potential solution)
- **Player A:** “Yeah that’s true. Then we wanna place bedrock ahead. Oh, but don’t we want to repeat that? One, two, three…” (Proposes specific solutions + Asks if others have suggestions)
- **Player B**: “And we have to move forward” (Proposes specific solutions)

2.3 Summary of Results

In Study 1, we coded the entire three-hour gameplay data based on the CPS model shown in Table 1. In Study 2, we randomly selected a 90-second segment out of each five-minute period for the 20-minute videos. Factor analyses indicated reasonable fit to our theorized model. Correlational analyses provided evidence on the orthogonality of the facets and their independence to individual differences in prior knowledge, intelligence and personality. Regression analyses indicated that the facets predicted both subjective and objective outcome measures controlling for several covariates. Overall, the results support the validity of our CPS model (see [6] for full details).

3 Automated CPS Modeling from Spoken Language

The next step was to automatically model the aforementioned CPS facets. Given the prevalence of verbal communication, we sought to model the data from speech analyzed at the utterance level. Accordingly, we used IBM Watson Speech to Text service to transcribe participants' audio using data from Study 2. Three research assistants were trained to code the resultant 11,163 utterances for evidence of the indicators from our CPS competency model. We aggregated the indicators to obtain binary codes for the presence or absence of each of the three core CPS facets per utterance.

3.1 Modeling Approach

We used Random Forest classifiers trained on the frequency counts of words and two-word phrases (bag of n-grams). Additionally, we investigated an alternate word coding method so that features would theoretically generalize to other domains. For this, we used the Linguistic Inquiry Word Count (LIWC) to count the proportion of words in an utterance that belong to each of 73 pre-defined categories (e.g., positive affect). Any non-zero LIWC categories (i.e. that category was present in the utterance) were added as uni-grams.

We used *team-level* ten-fold nested-cross validation where all the utterances for a given team were in the training set or testing set, but never both, which is important for team-level generalizability. Within each testing fold, the training set was again split into five folds, one of which was a validation fold for hyperparameter tuning. For each validation fold, a model was fit and scored using every combination of hyperparameters via a grid search. The accuracy scores for each parameter combination across the five validation folds were averaged, and the hyperparameters which resulted in the highest average accuracy were preserved. A model was then fit on the full training set using these best hyperparameters, and predictions were made on the test fold. These predictions were pooled over the ten test folds before final accuracy metrics were computed. We tuned four hyperparameters using this method: 1) whether to include unigrams or bigrams (n-grams only, not applicable for LIWC categories); 2) whether to use a pointwise mutual information (to filter phrases [8]) of 2 or 4 (bigrams only); 3) minimum document frequency of n-grams (0%, 1%, or 2%), and 4) training set balancing method (random undersampling, random oversampling, and synthetic minority oversampling technique). Class distributions for the validation and testing sets were left unchanged.

3.2 Summary of Results

Despite imperfect automatic speech recognition (word error rates of 45%), the n-gram models achieved AUROC (area under the receiver operating characteristic curve) scores of .85, .77, and .77 for construction of shared knowledge, negotiation/coordination, and maintaining team function, respectively (70%, 54%, and 54% improvement over chance). The LIWC-category models achieved similar scores of .82, .74, and .73 (64%, 48%, and 46% improvement over chance).

Next, we used linear mixed effects models to investigate the relationship between the three CPS facets and the following CPS outcome variables assessed at the individual level: posttest score, subjective perception of the team's performance, and subjective perception of the collaboration process (see [9]). To examine whether the human-coded and model-predicted scores yielded similar effects, we constructed separate models for each, resulting in 27 models (3 facets × 3 outcome variables × 3 sources [human vs. n-gram vs LIWC-category]). We averaged the expert-coded utterance scores and the model-predicted utterance-level probabilities for each participant for inclusion as predictors. We included each individual's total words spoken, ACT score, whether the individual knew his/her teammates, and whether the individual was assigned to interact with the environment as control variables (covariates). Team identity was included as a random factor (intercept only) to account for nesting of individuals within teams.

We found that n-gram and LIWC model-derived facet scores yielded similar coefficients to human-coded scores. Specifically, both model-derived scores of construction of shared knowledge positively predicted posttest scores ($b = .09$, $p < .05$ for n-grams) and $b = .08$, $p < .10$ for LIWC), which was similar to the human codes ($b = .11$, $p < .10$).

4 Closing-the Loop – Providing Feedback on CPS processes

We plan to embed the validated models into the collaborative environment to monitor and provide feedback on the unfolding CPS processes. For example, if maintaining team function is high but shared knowledge construction is low because one member is consistently dominating, then the system might display the following message: *"You all seem to be getting along great! But make sure that everyone on the team gets a chance to contribute solution ideas."* Alternatively, if team members are all generally contributing to the problem-solving efforts, but there are some issues with communication, specifically active listening since some members interrupt or talk over others, then the message could be: "*Everyone is contributing great solution ideas. Please make sure to listen to each other first before talking.*"

The precise intervention strategies, when to intervene (real-time or as a mid- or after-task review), how frequently to intervene, how to render the interventions, and the level of intervention (team level, individual level, or both) awaits design, testing, and refinement. Once a prototype is developed, we will conduct a controlled experiment to evaluate the efficacy of automated CPS feedback to an appropriate control condition. Our prediction is that the feedback-enabled system will yield to enhanced CPS outcomes, an exciting possibility that ushers forth a new generation of CPS environments that support real-time assessment and intelligent intervention.

5 Acknowledgments

This research was supported by the National Science Foundation (NSF DUE 1745442) and the Institute of Educational Sciences (IES R305A170432). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the funding agencies.

6 References

1. Griffin, P., B. McGaw, Care, E. Assessment and teaching of 21st century skills., New York: Springer. (2012).
2. Care, E., C. Scoular, Griffin, P. Assessment of collaborative problem solving in education environments. Applied Measurement in Education, **29**(4), pp. 250-264. (2016).
3. Organisation for Economic Co-operation and Development (OECD), PISA 2015 Collaborative Problem Solving Framework. (2015).
4. Roschelle, J.,Teasley, S.D. The construction of shared knowledge in collaborative problem solving, in Computer supported collaborative learning, C.O.M. (Ed.), Editor. Springer: Berlin. pp. 69-97. (1995).
5. Nelson, L.M. Collaborative problem solving, in Instructional design theories and models: A new paradigm of instructional theory, C.M. Reigeluth, Editor. Routledge: New York, NY. pp. 241-267. (1999).
6. Sun, C., V. Shute, A. Stewart, J. Yonehiro, N. Duran, D'Mello, S.K. Toward a Generalized Competency Model of Collaborative Problem Solving. (in review).
7. Ploetzner, R.,VanLehn, K. The acquisition of qualitative physics knowledge during textbook-based physics training. Cognition and Instruction, **15**(2). pp. 169-205. (1997).
8. Park, G., H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, M. Kosinski, D.J. Stillwell, L.H. Ungar, Seligman, M.E. Automatic personality assessment through social media language. Journal of Personality and Social Psychology, **108**(6): pp. 934-952. (2015).
9. Stewart, A., D'Mello, S.K. Connecting the Dots Towards Collaborative AIED: Linking Group Makeup to Process to Learning, in Proceedings of the 19th International Conference on Artificial Intelligence in Education (AIED'18). Springer. pp. 545-556. (2018).