Evaluating digital poetry: Insights from the CAT

Carolyn Lamb, Daniel G. Brown, Charles L.A. Clarke
University of Waterloo

Abstract

We test the Consensual Assessment Technique on recent digital poetry, using graduate students in Experimental Digital Media as judges. Our judges display good interrater agreement for the best and worst poems, but disagree on others. The CAT by itself may not be suitable for use on digital poetry; however, the behavior of quasi-expert judges when attempting this task gives us clues towards evidence-based, domain-specific evaluation. We discuss the role of product-base evaluation in digital poetry, and produce a set of desiderata based on our judges' written responses: Reaction, Meaning, Novelty, and Craft.

Introduction

Evaluation is a topic of contention in computational creativity (Jordanous 2013). While various means of evaluating creativity have been proposed (Colton 2008; Jordanous 2013; Ritchie 2007), we are unaware of any rigorous validity tests of these methods. Additionally, while creativity may be domain-specific (Baer 1998), there is little in the way of domain-specific testing for computational creativity, except for ad hoc questionnaires used by individual researchers.

The Consensual Assessment Technique (CAT) (Amabile 1983) is a well-known evaluation technique from psychology, which has not been used before on digital poetry. We use the CAT protocol to ask judges to rate different poems and explain their judgments. Judges display broad agreement on the best and worst poems. Their qualitative responses illuminate the qualities associated with creativity in digital poetry. By analyzing judges' written responses, we identify four major desiderata for digital poetry: Reaction, Meaning, Novelty, and Craft. Evaluating digital poetry may be more complicated than typical uses of the CAT (e.g. children's collages) due to the heterogeneity and experimental nature of digital poetry. Our four desiderata, combined with process-based evaluation, can be used in a more standardized, evidence-based evaluation of a complex field.

We believe that paying attention to the product-based evaluations of experts is a necessary step towards full computational creativity. A truly creative computer system will be able to produce work in a creative field that is taken seriously by others in the field. Process-based evaluation will tell us about the system's techniques, but only product-based evaluation by experts can tell us if the result is as valuable as intended.

Background

The Consensual Assessment Technique (Amabile 1983) is a method for evaluating human creativity. The idea is to assess creativity as a social phenomenon mediated by experts. A small group of expert judges (for example, visual artists to judge collages), give assessments of a group of artifacts. If the judges broadly agree in their judgments, then the assessment is considered valid. Importantly, judges are not told by what standards they should judge creativity, but are trusted to use their expert judgment.

The CAT and its reliability have been extensively studied. Baer and McKool (Baer and McKool 2009) summarize current best practices:

- Judges must possess expertise in the domain being judged; novice judges have poor interrater reliability. What constitutes expertise is a matter for debate, and can vary depending on medium. Skilled novices can have decent reliability (Kaufman, Baer, and Cole 2009), but some theoretical experts, such as psychologists, do not.
- Judges make their judgment independently, without consulting other judges.
- Judges review the artifacts blindly, without knowing framing information such as the author's identity.
- Judges are not told how to define creativity or asked to explain their ratings.
- Judges rate artifacts on a numerical scale with at least 3 points.
- Judges use the full scale. The most creative artifacts in the group should be at the top of the scale, and the least creative should be at the bottom.
- The number of judges varies from 2 to 40, with an average just over 10.
- Interrater reliability should be measured with Cronbach's coefficient alpha, the Spearman-Brown prediction formula, or the intraclass correlation method.
- An interrater reliability of 0.7 or higher is considered good. Expert judges generally achieve interrater reliabilities between 0.7 and 0.9.

- While the CAT was designed for a homogeneous group of subjects, it works in practice even when the artifacts were made under different conditions.

The CAT has become a gold standard for assessing human creativity. It has good reliability as long as the judges possess sufficient expertise. Obtaining experts is the major bottleneck in performing the CAT. However, the CAT is reliable even with relatively few (Baer and McKool 2009).

The CAT is mentioned frequently in computational creativity, but rarely implemented. Pearce and Wiggins use a modified CAT to assess chorale melodies (Pearce and Wiggins 2007). They ask their judges about the "success" of each melody rather than its creativity. More importantly, they give the judges explicit guidelines about what factors to consider when making their evaluations. Despite this, Pearce and Wiggins' judges achieve reasonable pairwise inter-judge consistency: mean $r(26) = 0.65$.

Because of the lack of framing information, the CAT can be used only to evaluate a creative product, not the process behind it. Product and Process are two of the four perspectives from which creativity can be judged (Jordanous 2015). We will return to this topic in our Discussion section.

Jordanous's SPECS model shares features with the CAT—particularly the use of expert raters to rank artifacts—but includes detailed training on theoretical sub-components of creativity (Jordanous 2013).

To our knowledge, CAT-related methods have not been used previously to assess computational poetry.

Method

For our study, we chose graduate students in the Waterloo's Experimental Digital Media program (XDM) as our judges. XDM includes digital poetry among a variety of other avant-garde, multimedia art forms, and students in the program actively practice producing such art. We judged XDM students likely to understand the demands of poetry as a genre and the challenges of generating poetry with a computer.

Judges were given a set of 30 poems, in randomized order. The poems are listed in Table 1. They evaluated each poem on a scale from 1 to 5, with 1 being "least creative" and 5 being "most creative". They produced these ratings without group discussion. Judges were told that some poems were written by computers, and others by humans, using computers; they were not told which of the poems were which.

The 30 poems, although not labeled as such, came from three different groups. Ten, group A, were poems in which we judged the authors were trying to create a relatively autonomous creative system; all but one of these poems were taken from published papers in the field of computational creativity. Another ten, group B, were poems in which the human exerted tighter artistic control (for example, by handcrafting templates), and the computer's role was relatively limited. The final ten poems, group C, were poems generated using specific source material which remained recognizable in the final product—either "found" poetry or modifications to a well-known human poem. All poems were published between 2010 and 2015. We presented the poems in plain text and without their titles. When the published work was a generator producing arbitrarily many poems (for example, a Twitter bot), we provided a single generated poem. In cases where the generated poem was excessively long, a 1-page excerpt was provided. While excerpting may bias judge responses to long poems, this is relatively unimportant to our analysis.

Once all 30 poems had been rated, we began the qualitative portion of the study. Each judge was asked to go back over the poems and, for at least 3 poems, write an explanation for their judgment. We did not present this request until the judges had made all 30 quantitative ratings and did not permit them to change their quantitative answers once the qualitative portion began.

We obtained a total of seven judges, which is within normal bounds for the CAT (Baer and McKool 2009). Participation took 1 hour.

We analyzed our data in three steps. First, we calculated the intraclass correlation between the seven judges. Second, we used the Kruskal Wallis test to see if there was a difference between ratings of poems from groups A, B, and C. Third, we used open coding to determine what major factors were used by our judges in their qualitative evaluations.

Results

Interrater reliability

The intraclass correlation between our judges (a statistic that can range from -1.0 to 1.0) was only 0.18—far below the 0.7 to 0.9 standard for CAT results. A bootstrapped sample of 10,000 permuted versions of the data showed that interrater agreement hovered around zero, with a standard deviation of 0.04, meaning that if each judge gave their ratings by chance, there would be much less correlation between ratings than what our results show. The agreement between our judges, therefore, is statistically significant at $p < 0.01$. However, it is not a strong enough agreement to be used for the usual applications of the CAT, such as judging admissions to academic programs in creative fields.

Looking at the data, some poems were rated very highly by nearly all judges, while others were rated very poorly, but a large mass of poems in the middle had inconsistent or inconclusive results. When the data is reduced to those poems with the highest and lowest average scores, intraclass correlation becomes very high. With the seven best and seven worst poems—nearly half of our original data set—the intraclass correlation is 0.73, and narrowing the number of poems raises that statistic still higher. It is intuitive that poems with the highest and lowest ratings would have relatively good agreement, while poems about which judges disagreed would have average scores closer to the middle. However, when we looked at only the seven best and seven worst in each of our 10,000 bootstrap samples, the mean and standard deviation for intraclass correlation did not rise. Therefore, the agreement on the best and worst poems is not merely a statistical artifact; our judges really were able to agree on these ends of the spectrum.

It appears that our judges agree when selecting the best and worst poems in a group, but cannot reliably rank the group of poems as a whole.

Group	Title	Author	Score	Response
B	“Notes on the Voyage of Owl and Girl”	J.R. Carpenter	4.4	6
B	Excerpt from “Definitions II - Adjectives”	Allison Parrish	4.3	18
C	“Conditionals”	Allison Parrish	4.0	5
B	“trans.mission [a.dialogue]”	J.R. Carpenter	3.9	4
A	Untitled	Unnamed system (Toivanen et al. 2013)	3.6	0
B	“Walks From City Bus Routes”	J.R. Carpenter	3.6	11
B	Excerpt from “]]”	Eric Goddard-Scovel and Gnoetry	3.4	0
C	“St. Louis Blues 2011”	Christopher Funkhouser	3.3	10
A	Untitled	P.O. Eticus (Toivanen, Gross, and Toivonen 2014)	3.1	0
A	“Angry poem about the end”	Unnamed system (Misztal and Indurkhya 2014)	3.1	0
B	“Good Sleep”	George Trialonis and Gnoetry	3.1	10
A	“Voicing an Autobot”	Allison Parrish	2.9	4
B	“The Ephemerides”	Allison Parrish	2.9	0
A	Untitled	IDyOt (Hayes and Wiggins)	2.9	3
C	“HaikU”	Nanette Wylde	2.8	5
C	Excerpt from “Dark Side of the Wall”	Bob Bonsall	2.8	0
B	Excerpt from “Exit Ducky?”	Christopher Funkhouser, James Bonnici, and Sonny Rae Tempest	2.7	9
C	“Spine Sonnets”	Jody Zellen	2.6	1
C	“Ezra Pound Sign”	Mark Sample	2.6	13
A	“Blue overalls”	Full-FACE (Colton, Goodwin, and Veale 2012)	2.6	5
C	“Regis Clones (Couplets from ZZT-OOP)”	Allison Parrish	2.6	6
A	“quiet”	MASTER (Kirke and Miranda 2013)	2.4	11
B	Excerpt from “Permutant”	Zach Whalen	2.2	12
A	Untitled limerick	Unnamed system (Rahman and Manurung 2011)	2.2	3
A	“The legalized regime of this marriage”	Stereotrope (Veale 2013)	2.1	7
C	“Times Haiku”	Jacob Harris	2.1	10
C	Untitled	Mobtwit (Hartlová and Nack 2013)	2.1	15
A	Untitled	Unnamed system (Tobing and Manurung 2015)	1.9	4
B	“Rapbot”	Darius Kazemi	1.9	0
C	“The Longest Poem in the World”	Andrei Gheorghe	1.7	7

Table 1: The 30 poems used in our experiment, ranked from highest to lowest average rating. The “Response” column lists how many lines of explanation, in total, were given for judges’ ratings of the poem in part 2 of the study.

Calculating Kendall’s tau between pairs of judges did not result in any useful clusterings of the judges into factions.

Kruskal-Wallis test

While a disproportionate number of the best-ranked poems were from Group B, a Kruskal-Wallis test showed that this difference was not significant ($\chi^2_2 = 3.8$, $0.25 > p > 0.1$). Both relatively creative and relatively uncreative poems existed in groups A, B, and C, and no group did systematically better than others.

The individual poems, their group membership, and their average scores are shown in Table 1.

Open coding of qualitative data

We performed open coding by giving each line of written response a content label and a valence (positive, negative, or neutral), then clustering the content labels into categories. Each category gave insight into the implicit values used by our judges. Overall there were 179 coded lines distributed over 63 explained judgments—an average of 2.8 lines per judgment and 9 judgments per reader. Table 2 shows the proportions of lines of each type and their valences.

One coder performed the initial clustering, while a second repeated the labeling to validate the first coder’s responses. Our two coders agreed on roughly half of lines as to the exact categorization, and for two thirds of lines agreed as to the general category. For the remaining one third, in half of cases, the reviewers agreed that a line could easily be coded as having both reported categories, such as clich imagery, which has to do with both novelty and imagery. Of the remaining ones, the coders did not initially agree, but were quickly convinced of one or the other categorization.

Grouping	Category	% of comments	% Positive	% Negative
Reaction	Feeling	12%	68%	31%
Reaction	Comparison	10%	44%	44%
Reaction	Base/Other	12%	57%	29%
Meaning	Message	14%	60%	40%
Meaning	Coherence	7%	67%	33%
Meaning	Content	5%	89%	0%
Craft	Technique	12%	67%	29%
Craft	Imagery	7%	67%	33%
Craft	Form	4%	50%	50%
Craft	Skill	3%	20%	80%
Novelty	Novelty	15%	42%	58%

Table 2: Categories derived from our qualitative data.

Below we explain the meanings of each of our labels.

Reaction. 34% of lines described, structured, or contextualized the judge's affective reaction to the poem.

Feeling. Statements about the emotions evoked in the judge by the poem.

- "This was just super fun to read."
- "Felt empty."

Comparison. Lines that compared the poem to something else, including existing poems or poetry movements.

- "I like how this echoes, say, Siri giving instructions."
- "This reminds me of bad, early 2000s my space poetry... Angsty teens spewing 'creativity' on the world wide web."

Base/Other. Base lines are statements that the poem is or is not creative, without immediate explanation. Often a judgment containing Base lines contained explanation in other lines, so a Base line can be thought of as a topic sentence, not necessarily an unsupported judgment. Lines coded "Other" similarly contain statements more to do with structuring a judgment than with the judgment itself, such as statements that the judge felt conflicted about the poem.

Meaning. 26% of lines described the meaning of the poem, the concepts involved, and their clarity.

Message. Statements about the idea that the judge believes the poet intended to communicate. Most judges made comments with negative valence, stating that a poem was not sufficiently meaningful. A major exception was judge 7, who left long positive comments closely interpreting the meaning of several poems.

- "It would be more creative/interesting if there were a distinct theme or repetition of some sort—some sort of message to the reader."
- "It begins with an opinion about a campy TV show and ends on gleeful nihilism. The real American Horror Story is the nuclear apocalypse, the end of the world effected by some hideous war games between two self-obsessed nations flexing their muscles at each other. (good twerking). It packs a lot into a very compressed collection of sentences, and also manages to serve as a brutal indictment of contemprary culture."

Coherence. Nearly all of the lines we coded as Coherence referenced a lack of coherence, or nonsense.

- "It felt too disjointed."

Content. Statements about the characters, objects, or events in the poem. Judges mentioned this aspect of meaning less frequently than more abstract ideas.

- "It's a complete narrative in just 3 very short lines."

Craft. 20% of lines described the way in which the poem's concept was executed.

Technique. Statements assessing specific literary techniques used in the poem. They include defamiliarization, enjambment, phrasing, repetition, rhyme, rhythm, vocabulary, and voice, as well as more general statements such as "playful use of language". Poor technique, as displayed by a limited vocabulary or by the poem seeming "forced", was coded negative.

- "I like this one because I feel like I can hear a distinct voice."
- "I found the creative intentions - caps, quotation marks, the fragmentive narrative, the asterisks - forced and not really used well."

Imagery. Statements commenting on the poem's use of imagery. Imagery is a specific type of content involving direct sensory descriptions, and is important in contemporary poetry (Kao and Jurafsky 2012).

- "Good consistent imagery and figurative language."
- "The imagery isn't provocative."

Form. Only a few poems received comments on their form. Three poems in inventive forms, such as imaginary dictionary entries, were praised for these concepts. Another received a comment that it was too short. In addition, two haikus received negative comments for lacking subtle features of the traditional haiku.

- "If there were another stanza I'd like it more."

Skill. Statements assessing the poet's skill or cleverness.

- "Very rudimentary and woe is me."

Novelty. 15% of lines were statements about the poem's novelty. Positive valence lines stated that the poem was unusual, unique, or subversive. Negative valence lines stated that the poem—or aspects of the poem—were obvious, derivative, unoriginal, trite, clichéd, banal, failed to push boundaries, or did not sufficiently change their source text.

- "I don't think this is very creative because it doesn't push the boundary of poetry in any way."
- "This is creative because its unique. I've never seen a poem like this before."

Judges frequently disagreed on what traits a poem possessed, and on the valence assigned to those traits. A poem might be described as incoherent by one judge but interestingly disjointed by another, or banal by one judge but unexpected by another. An extreme example is a poem generated by Mobtwit (Hartlová and Nack 2013). The poem was written by arranging tweets to generate emotional contrast. It was described as random and devoid of meaning by Judge 1 (who rated the poem as 1 out of 5), but Judge 7 rated the poem a 5 and gave a long exposition of its meaning (quoted above, under "message"). Restricting the sample to the seven highest and seven lowest rated poems did not remove these qualitative disagreements.

There were slightly more positive (93) than negative (79) lines overall. There was a modest positive correlation between quantitative score and number of positive comments, a similar modest negative correlation between quantitative score and number of negative comments, and no correlation at all between quantitative score and total comments ($r = 0.26$, -0.27, and 0.008 respectively).

Discussion

Since we did not achieve the usual inter-rater reliability standard of the CAT, our method is not a finished evaluation. It is possible that the CAT will not provide standardized computational poetry evaluation at all. However, the qualitative portions of our study illuminate how judges with some expertise evaluate computational art, which leads us to a better understanding of what criteria could go into such an evaluation in the future.

Judge selection

Why did our judges disagree about poems in the middle of the set? Should we have chosen a different set of judges? We believe that our judges' lack of interrater reliability speaks to something more complex than a simple lack of expertise.

The question of who, exactly, has sufficient expertise for the CAT is a difficult one. Kaufman et al. review prior work in the differences between expert CAT judges and novices (Kaufman, Baer, and Cole 2009) . Novices lack the interrater reliability of experts and their judgments only moderately correlate with expert judges. However, in many cases, gifted novices (which Kaufman et al. describe as "quasi-experts") produce judgments that are more in line with those of experts than with the general population. Novices have fewer problems serving as judges when the art form in question is one that the general population encounters regularly: stories rather than poems, for instance. However, psychologists—even psychologists of creativity—are not experts; they perform as inconsistently as novices from the general population. The expertise necessary for the CAT seems to have more to do with experience in a specific creative field than with knowledge of the theoretics of creativity.

Pearce and Wiggins used both music researchers and music students as judges (Pearce and Wiggins 2007). Why did their experiment achieve close to the recommended interrater reliability while ours did not? One answer is that Pearce and Wiggin's study was an evaluation of chorale melodies, which are simpler, less diverse, and defined by more well-established rules than computational poetry.

We argue that Experimental Digital Media students should be considered quasi-experts. Even more advanced than Kaufman et al.'s gifted novices, these students are more like experts-in-training, undergoing advanced education in how to produce art in their field. However, the field of digital poetry is too new to be well-defined. It is also possible that the different poets in our study are performing different tasks that ought not to be grouped together. The CAT's more typical uses revolve around homogeneous products, such as the poetry or collages of elementary school students. Mature artists and researchers, in a new field where a variety of movements, motivations, and techniques are still under development, likely produce a more complex and contentious body of work.

A good idea for future work might be to replicate the CAT with other groups of experts and quasi-experts, or with a more homogeneous group of digital poems. Poets who have been paid for their published work, or participants in events such as the E-Poetry Festival (Glazier 2016), might be appropriate experts.

However, we strongly advise against the use of computational creativity researchers as expert judges unless they themselves are practicing artists in the field being studied. Computer researchers without such artistic experience are likely to have the same problem as psychologists judging human art. They may thoroughly understand the theory, but they are unlikely to have an expert sense of the *artistic* aspects of their work. Moreover, because academic publishing depends heavily on theory and argumentation, and because the field of computational creativity is so new, computer researchers (including ourselves) are likely to be distracted from evaluations of specific products by our beliefs about where we would like the field to go.

Reliance on experts

As noted, novices lack high interrater reliability, and their judgments correlate only modestly with those of efforts. In some areas, novice judgments can be uncorrelated or even negatively correlated with those of experts (Lamb, Brown, and Clarke 2015). However, some researchers have good reasons for setting a goal of popular appeal rather than the approval of experts. For these groups, techniques based on interrater reliability are not suitable, since novices lack it and experts are not the intended audience. Popular appeal should be measured through other methods, such as perhaps Jor-

danous's measurements of community impact (Jordanous, Allington, and Dueck 2015).

Judge bias

Specific to computational creativity is the possibility of judges being biased against computational art, due to pre-existing beliefs about what computers can and can't do, or to a need to connect with the imagined human author. Some researchers suggest providing framing information in order to fix this problem (Charnley, Pease, and Colton 2012). However, this bias does not always empirically appear.

Friedman and Taylor told judges either that musical pieces were composed and performed by humans or that they were composed and performed by computers. Judges' beliefs about who composed the music did not significantly moderate their enjoyment, emotional response, or interest in the music (Friedman and Taylor 2014). This was true regardless of the judges' expertise. Similarly, Norton et al. found that while individual humans can be biased for or against computers, the bias across a group was usually not statistically significant (Norton, Heath, and Ventura 2015).

Our anecdotal experience suggests that XDM students are in little danger of bias against computers. They themselves incorporate computers into their process on a daily basis. If anything the bias was in the other direction. As one judge put it after the experiment, "Sometimes I wanted to say that a poem was childish, but then I thought, 'What if a computer wrote it?' and I didn't want to hurt the computer's feelings."

Even where bias against computers exists, it is not relevant unless computer products are compared with the products of humans and the judges are somehow aware of which products are from which group. All the poems in our study had some involvement from both humans (who wrote a computer program) and computers (which put together words based on the program), but judges were not told what the computer's role was. It is easy to imagine studies where the role of the computer is more homogeneous, or even studies which compare outputs from different versions of a single program.

Desiderata for domain-specific poetry evaluation

Baer argues that creativity is an umbrella term for a variety of independent domain-specific skills (Baer 1998). If this is the case, then evaluations of computational poetry would be expected to contain criteria that apply only to poetry, perhaps only to computational poetry. Studies like ours are a step towards developing these criteria.

Our study suggests a set of desiderata shared by most of our judges for poetry:

- **Reaction.** The poem should provoke feelings of enjoyment and/or interest from the reader.
- **Meaning.** The poem should intentionally convey a specific idea. Even if the poem is difficult to understand, its difficulty should enhance the underlying meaning. (For example, a Dadaist poem uses apparently meaningless text to illustrate ideas about how language and meaning work.)
- **Novelty.** The poem should be unusual or surprising in some way, and not merely repeat familiar tropes.
- **Craft.** The poem should make effective use of poetic techniques in service of the other three criteria. This can include form, imagery, auditory effects such as rhyme, psychological effects such as defamiliarization, visual effects such as enjambment, and verbal effects such as voice. Effective use of these techniques requires skill.

These desiderata are not straightforward. In particular, some of the literary techniques praised by our judges oppose each other. At least one poem received positive comments for its detailed imagery, while another received a positive comment for simplicity. Requiring detail and simplicity at the same time is a contradiction!

We suggest viewing literary techniques as a toolbox of strategies for poetic success. Some may be more appropriate to a particular goal than others. The question asked to a judge about craft should not be, "How many times are literary techniques used?" It should be something more like, "What techniques are used, and how effective are they?" It should be assumed that such questions can only be answered by expert or quasi-expert judges.

Relations between our desiderata and existing theories

Our desiderata have overlap with other evaluation theories, but are not identical to them. For example, Novelty and Value are frequently used to evaluate creativity. Our judges did emphasize Novelty, but Value either did not appear or was divided into many sub-criteria.

Van der Velde et al. use word association to define creativity criteria (van der Velde et al. 2015). Our Novelty criterion corresponds to their Original and Novelty/Innovation, while their Skill and Craftsmanship correspond to our Craft. Van der Velde's other criteria are Emotion and Intelligence.

Ritchie suggests Typicality and Quality criteria (Ritchie 2007). A hint of Typicality can be seen in Comparison judgments. It appears that for a positive typicality judgment, a poem must strike the judge as not merely typical of poetry, but typical of *good* poetry. The "Dadaist dictionary" was rated highly, but poems typical of "the scrawl of a high school senior" were not. Groundedness in relevant poetic movements led to a positive response, but so did poems seen as entirely novel. Typicality as Ritchie conceives it may be neither necessary nor sufficient for computational poetry.

The Creative Tripod (Colton 2008) consists of Skill, Imagination, and Appreciation. While Skill as such was a minor category for us, everything under the Craft grouping presumably requires skill. Imagination was rarely mentioned, but it could be argued, as by Smith et al., that Imagination is the underlying trait which allows for Novelty (Smith, Hintze, and Ventura 2014). This reading is supported by Van der Velde et al., who group "Imagination" under Novelty/Innovation (van der Velde et al. 2015). Appreciation is difficult to read into any of the coded comments. However, the highly fluid definitions of traits in the Tripod make it difficult to definitively state if they are present or not.

Manurung et al.'s criteria of Meaningfulness, Poeticness, and Grammaticality (Manurung, Ritchie, and Thompson 2012) overlap with our desiderata. Meaningfulness and Meaning are synonymous; Poeticness and Craft are similar concepts. However, Manurung et al. operationalize Poeticness as meter and rhyme, while judges in our study had a more expansive view of Craft. Grammaticality was not emphasized; some poems received positive comments despite being quite ungrammatical.

Our Reaction criterion does not appear in many existing models, since most models focus only on qualities imputed to the poem or poet. However, it bears some resemblance to the Wellbeing and Cognitive Effort criteria of the IDEA model (Colton, Pease, and Charnley 2011), which could perhaps be used to break judge reactions down more finely.

The product or the process?

We believe that product-based evaluation is an important part of the creative process. Nevertheless, it has a major drawback: it cannot differentiate between the creativity of the computer system and the creativity of the human who programmed it.

Some examples from our data set illustrate this problem. "Notes on the Voyage of Owl and Girl", our most highly rated poem, is based on a tightly handcrafted template. The human author provides a narrative structure which does not alter, and the computer selects details (from a human-curated list) to fill it in. In its original form, "Owl and Girl" exists on a web page and is periodically re-generated before the viewer's eyes. "Owl and Girl" is interesting artistically, but its high ratings refer mostly to the creativity of the human author.

Conversely, "The legalized regime of this marriage", created by Stereotrope, is among the most poorly rated. Stereotrope is an experiment in computational linguistics. The system mines existing text for similes, produces a common-sense knowledge base using these similes, and uses the knowledge base to generate similes and metaphors of its own. The new similes and metaphors are then used to fill in templates and construct a poem. Our judges disliked this poem, calling it obvious, clichéd, unskilled, and uncreative. However, Stereotrope is doing something more *computationally* interesting than "Owl and Girl".

We must ask what the goal is for a system like Stereotrope. Do we wish to construct a system whose use of simile and metaphor is artistically successful? Then Stereotrope—in its current form—fails. But if we wish to construct a system using *humanlike* simile and metaphor, then it is easy to argue that Stereotrope succeeds: its metaphors feel obvious *because* they are humanlike. Such a system might not be artistically creative, but it might be a good model of the everyday creativity of non-artist humans expressing themselves. We will not know if a system has succeeded unless we know which of these goals it was aiming for. (Other goals than these are, of course, possible.)

We believe that ultimately, computational creativity must succeed on both fronts. To set a goal of artistic success while ignoring process is to abandon comparison to human creativity. But to set a goal of process while ignoring product is to fail to take seriously the very medium in which the computer is working. A system which fails to take art seriously can have value as a cognitive model, but that model will not represent the cognition of skilled human artists, nor will the system's output be taken seriously by such artists.

One could argue that a computer must first establish a humanlike process before refining that process to be more artistic. This is reasonable, but debatable. It is also possible that producing good art and using a humanlike process are two tasks at which the computer can progress simultaneously. The learning process of an initially-uncreative computer may or may not look like the learning process of an initially-unskilled human, and setting a goal of behaving like an unskilled human may in some circumstances be counterproductive.

An interesting idea for future work would be to replicate the CAT study and present information about the specific tasks assigned to the computer, in a standardized form such as the diagrams in (Colton et al. 2014). CAT judges would then be asked how creative they believed *the computer* had been. An alternative would be to use one evaluation technique for product, and another for process.

Conclusion

The Consensual Assessment Technique is an established product-based creativity evaluation. We used the CAT to examine the opinions of Experimental Digital Media students, whom we consider quasi-experts, on computational poetry. The students agreed on the best and worst poems, but were divided about the ones in the middle. There was no significant bias for or against poems from the computational creativity research community.

Based on qualitative comments, we identified several evidence-based criteria through which our judges made their evaluative decisions: Reaction, Meaning, Novelty, and Craft. These might be refined for use in future evaluations.

Despite modest overall inter-rater reliability, we argue that Experimental Digital Media students are appropriate quasi-experts. Achieving consistency may be difficult for computational poetry due to its experimental and diverse nature. It is important for evaluation to take into account both product and process. Our present study does not include process components, but process could be added, either separately, or by including process information in some systematic way. We believe there is more knowledge to be gained by investigating the workings of the CAT and other expert judgment procedures.

References

Amabile, T. M. 1983. A consensual technique for creativity assessment. In *The Social Psychology of Creativity*. Springer. 37–63.

Baer, J., and McKool, S. S. 2009. Assessing creativity using the consensual assessment technique. *Handbook of Assessment Technologies, Methods and Applications in Higher Education* 65–77.

Baer, J. 1998. The case for domain specificity of creativity. *Creativity Research Journal* 11(2):173–177.

Charnley, J.; Pease, A.; and Colton, S. 2012. On the notion of framing in computational creativity. In *Proceedings of the Third International Conference on Computational Creativity*, 77–82.

Colton, S.; Pease, A.; Corneli, J.; Cook, M.; and Llano, T. 2014. Assessing progress in building autonomously creative systems. In *Proceedings of the Fifth International Conference on Computational Creativity*, 137–145.

Colton, S.; Goodwin, J.; and Veale, T. 2012. Full face poetry generation. In *Proceedings of the Third International Conference on Computational Creativity*, 95–102.

Colton, S.; Pease, A.; and Charnley, J. 2011. Computational creativity theory: The face and idea descriptive models. In *Proceedings of the Second International Conference on Computational Creativity*, 90–95.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *AAAI Spring Symposium: Creative Intelligent Systems*, 14–20.

Friedman, R. S., and Taylor, C. L. 2014. Exploring emotional responses to computationally-created music. *Psychology of Aesthetics, Creativity, and the Arts* 8(1):87–95.

Glazier, L. P. 2016. E-poetry: An international digital poetry festival. http://epc.buffalo.edu/e-poetry/archive/, accessed February 25, 2016.

Hartlová, E., and Nack, F. 2013. Mobile social poetry with Tweets. Bachelor thesis, University of Amsterdam.

Hayes, M. D., and Wiggins, G. A. Adding semantics to statistical generation for poetic creativity. Late-breaking abstract at the *Sixth International Conference on Computational Creativity*, 2015.

Jordanous, A.; Allington, D.; and Dueck, B. 2015. Measuring cultural value using social network analysis: a case study on valuing electronic musicians. In *Proceedings of the Sixth International Conference on Computational Creativity June*, 110.

Jordanous, A. K. 2013. *Evaluating Computational Creativity: a standardised procedure for evaluating creative systems and its application*. Ph.D. Dissertation, University of Sussex.

Jordanous, A. 2015. Four PPPPerspectives on Computational Creativity. In *Proceedings of the AISB Symposium on Computational Creativity*. 8 pages.

Kao, J., and Jurafsky, D. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *NAACL Workshop on Computational Linguistics for Literature*, 8–17.

Kaufman, J. C.; Baer, J.; and Cole, J. C. 2009. Expertise, domains, and the consensual assessment technique. *The Journal of Creative Behavior* 43(4):223–233.

Kirke, A., and Miranda, E. 2013. Emotional and multiagent systems in computer-aided writing and poetry. In *Proceedings of the Artificial Intelligence and Poetry Symposium (AISB13)*, 17–22.

Lamb, C.; Brown, D. G.; and Clarke, C. L. 2015. Human competence in creativity evaluation. In *Proceedings of the Sixth International Conference on Computational Creativity June*, 102.

Manurung, R.; Ritchie, G.; and Thompson, H. 2012. Using genetic algorithms to create meaningful poetic text. *Journal of Experimental & Theoretical Artificial Intelligence* 24(1):43–64.

Misztal, J., and Indurkhya, B. 2014. Poetry generation system with an emotional personality. In *Proceedings of the Fifth International Conference on Computational Creativity*, 72–81.

Norton, D.; Heath, D.; and Ventura, D. 2015. Accounting for bias in the evaluation of creative computational systems: An assessment of DARCI. In *Proceedings of the Sixth International Conference on Computational Creativity*, 31–38.

Pearce, M. T., and Wiggins, G. A. 2007. Evaluating cognitive models of musical composition. In *Proceedings of the Fourth International Joint Workshop on Computational Creativity*, 73–80.

Rahman, F., and Manurung, R. 2011. Multiobjective optimization for meaningful metrical poetry. In *Proceedings of the Second International Conference on Computational Creativity*, 4–9.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Smith, M. R.; Hintze, R. S.; and Ventura, D. 2014. Nehovah: A neologism creator nomen ipsum. In *Proceedings of the International Conference on Computational Creativity*, 173–181.

Tobing, B. C., and Manurung, R. 2015. A chart generation system for topical metrical poetry. In *Proceedings of the Sixth International Conference on Computational Creativity June*, 308–314.

Toivanen, J. M.; Järvisalo, M.; Toivonen, H.; et al. 2013. Harnessing constraint programming for poetry composition. In *Proceedings of the Fourth International Conference on Computational Creativity*, 160–167.

Toivanen, J. M.; Gross, O.; and Toivonen, H. 2014. The officer is taller than you, who race yourself! Using document specific word associations in poetry generation. In *Proceedings of the Fifth International Conference on Computational Creativity*, 355–359.

van der Velde, F.; Wolf, R. A.; Schmettow, M.; and Nazareth, D. S. 2015. A semantic map for evaluating creativity. In *Proceedings of the Sixth International Conference on Computational Creativity June*, 94.

Veale, T. 2013. Less rhyme, more reason: Knowledge-based poetry generation with feeling, insight and wit. In *Proceedings of the Fourth International Conference on Computational Creativity*, 152–159.