# Human Competence in Creativity Evaluation

**Carolyn Lamb, Daniel G. Brown, Charles L.A. Clarke**
School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
c2lamb@uwaterloo.ca, dan.brown@uwaterloo.ca, claclark@plg.uwaterloo.ca

## Abstract

We investigate the performance of non-expert judges in using leading computational poetry evaluation metrics to evaluate poetry written by humans. We find that regardless of the model used, non-expert judges are very poor at using metrics to evaluate creativity, even displaying the reverse of the desired rating pattern, preferring novice poetry to professional poetry. We discuss likely reasons for this finding and the implications for the evaluation of computational creativity. Researchers using human judges should be aware that using a metric or structured evaluation does not negate the need for judge expertise.

## Introduction

An increasingly important debate in Computational Creativity is the development of standardised evaluation methods. There are many reasons why it is desirable for computers to recognize and evaluate creativity, including the assistance of humans in creative acts, understanding of the creative human mind, and the AI application of teaching computers to behave creatively themselves. However, it is not clear how exactly one would go about distinguishing more creative from less creative output. Two important camps in this debate are those who use a metric with specific criteria (e.g. (Pease, Winterstein, and Colton 2001; Ritchie 2007; Colton 2008a; Colton, Pease, and Charnley 2011) and those who prefer a consensual assessment based on the agreement of expert judges, without specific criteria (Amabile 1983).

While the Consensual Assessment Technique has been rigorously tested (see e.g. (Kaufman, Baer, and Cole 2009)), specific metrics used in the field of Computational Creativity have not. We therefore undertook an empirical test of four such metrics from the existing literature. These metrics evaluate a product's creativity based on (for example) its novelty, value, skill and other qualities, or on some calculation involving these qualities.

We collected poems generated by humans at various levels of skill. We then recruited a large number of humans to evaluate the poems on the criteria used in our selected metrics. Our results were very counter-intuitive. On nearly every criterion, our judges significantly preferred amateur, unskilled poems to the work of professional poets—the reverse of what one would expect.

Poetry is a rarefied field, and we suspected that the reversed results were caused by untrained raters having difficulty understanding the professional poems. Such poetry might not be accessible to an untrained reader. We ran the experiment again with poems written for children. This second experiment did not produce reversed results, but any power of the criteria to differentiate between good and bad poetry was reduced to noise.

Our experiments show that non-expert judges do not apply creativity metrics appropriately to poetry. Of course, the Consensual Assessment Technique already mandates the use of expert judges for this reason. Non-experts in a consensual assessment have poor inter-rater reliability and poor agreement with the judgments of experts (Kaufman, Baer, and Cole 2009). However, our research shows that this problem also applies to judgments made with specific criteria. Using such criteria is not an escape from the issue of judge selection. Moreover, beyond simply losing reliability, the use of non-expert judges can produce the exact opposite of the intended result.

Many evaluations in computational creativity today are still done by the researchers themselves (Colton, Goodwin, and Veale 2012; Norton, Heath, and Ventura 2010; Riedl and Young 2006; Smith, Hintze, and Ventura 2014) or by a group of human volunteers whose expertise in creativity is not discussed (Burns 2015; Gervás 2002; Karampiperis, Koukourikos, and Koliopoulou 2014; Llano et al. 2014; Monteith, Martinez, and Ventura 2010; Norton, Heath, and Ventura 2013; Román and y Pérez 2014). For robust evaluation, it may turn out that neither of these approaches is sufficient.

## Background and Related Work

The past 25 years of computational creativity research owe much to Boden's (Boden 1990) work on the meaning of creativity. Boden focuses on creativity as the exploration and transformation of conceptual space. While Boden's book does not give a definition which can be broken down into formulaic parts, she does repeatedly mention the need for creative systems to produce works which are both novel and valuable. Subsequent researchers have built on her work to propose numerical metrics.

Ritchie, the first such researcher, proposes that human creativity is evaluated according to the criteria of Novelty

("To what extent is the produced item dissimilar to existing examples of that genre?") and Quality ("To what extent is the produced item a high-quality example of that genre?") (Ritchie 2001). For computational creativity, he proposes replacing Novelty with Typicality—as a computer program must first be able to generate plausible examples of a type of creative product before attempting to make ones dissimilar from what has gone before. Ritchie then suggests various tentative criteria, such as "high quality items should make up a significant proportion of the results", for evaluating a system based on its Typicality and Quality over several runs. The presence of these composite criteria implies that using Typicality and Quality measurements directly for creativity evaluation, without further analysis, may be overly simplistic. Nevertheless, one can easily imagine common-sense constraints on the base measurements. For example, while the Quality measurement could be used in various ways, one would certainly not expect creative poems to have a lower average Quality than uncreative ones.

Ritchie's model has been used to evaluate creative systems in practice (e.g. (Gervás 2002; Tearse, Mawhorter, and Wardrip-Fruin 2011)). Other researchers performing similar work focus on Novelty rather than Typicality, a choice more in line with Boden's work. For example, Pease *et al.* (Pease, Winterstein, and Colton 2001) suggest a variety of ways to formally measure both Novelty and Value (a synonym of Quality).

Some difficulties in the Boden-based models, particularly Ritchie's, have been illuminated through experience. Many of Ritchie's composite criteria are based on comparisons with an inspiring set of existing work. In the absence of a quantitative measure for similarity between creative products, such criteria are difficult to evaluate (Gervás 2002). Ventura's RASTER thought experiment (Ventura 2008) also claims to illustrate flaws in Ritchie's model: a highly uncreative system, generating works completely at random, can technically be said to meet the criteria. However, the RASTER thought experiment uses images from a Web search to guide output, without considering those images an inspiring set. It also fails to consider typicality and quality independently, which renders many criteria inapplicable. Ventura suggests that the inapplicability of these criteria, in and of itself, is a reason to treat a system with suspicion.

Another metric, Colton's Creative Tripod (Colton 2008a), judges creative work by whether it appears to be skillful, appreciative, and imaginative. Colton's tripod has frequently been used to evaluate creative systems (Smith, Hintze, and Ventura 2014; Chan and Ventura 2008; Monteith, Martinez, and Ventura 2010; Young, Bown, and others 2010) or to guide their development (Norton, Heath, and Ventura 2010; Colton 2008b). A weakness of the tripod is that specific definitions for the three criteria are not provided. It has been pointed out (Bown 2014) that this provides too much opportunity for authors to make impressionistic statements about why their system meets the criteria, without rigorous, falsifiable inquiry into whether its performance in these areas is sufficient. Even the intentionally uncreative RASTER (Ventura 2008) is argued to meet Colton's criteria in this manner.

Colton *et al.* have added many words to the tripod since its construction, including Learning, Intentionality, Accountability, Innovation, Subjectivity, and Reflection (Colton et al. 2014). However, since the majority of recent work implementing the tripod uses only the original three words, we focus our research on these original three.

Another proposal by Colton *et al.* is the IDEA model (Colton, Pease, and Charnley 2011), in which an ideal audience rates a creative product according to Wellbeing (how much they likes the product) and Cognitive Effort (how prepared they are to spend effort thinking about and interpreting it). Like the criteria of Ritchie's model, Wellbeing and Cognitive Effort can be combined to measure different aspects of a product's reception. For example, if the variance in Wellbeing is high, a product would get a high score on "Divisiveness".

Many other standardized metrics for evaluating a creative system have been proposed. Jordanous's SPECS model (Jordanous 2012) incorporates many criteria based on cultural beliefs about the meaning of creativity, including criteria similar to Novelty and Value. Burns's EVE' model defines creativity as a combination of Surprise and Meaning, and has been applied to humorous poetic advertisements (Burns 2015), humorous haiku (Burns 2012) and, in thought experiment form, to line drawings (Burns 2006). Other new metrics either proposed or used ad hoc in the past ten years come from varied sources including Piaget's theories of cognitive development (Aguilar and Pérez y Pérez 2014), theories about quality in a specific art form (Das and Gambäck 2014; Rashel and Manurung 2014; Pearce and Wiggins 2007), interestingness (Román and y Pérez 2014; Gervás 2007), and many others (Brown 2009; Lehman and Stanley 2012; Llano et al. 2014; Monteith et al. 2013; Norton, Heath, and Ventura 2013).

Very rarely have any such metrics been validated through direct use on human-generated products. A few researchers have used the metrics to compare computational products to human-generated products. Monteith *et al.* compare human-composed to computer-composed music using an operationalization of Colton's tripod (Monteith, Martinez, and Ventura 2010). The computer music did better at expressing specific emotions (Skill) but the human music sounded more like "real music" (Appreciation). Burns tested his EVE' model on human products (Burns 2015) and found good correspondence between his model and human ratings; Surprise multiplied with Meaning accounted for 70% of the variability in ratings of Creativity.

Binsted *et al.* built a system, JAPE, to generate riddles (Binsted, Pain, and Ritchie 1997), and evaluated it using children's responses to criteria similar to those which would later form Ritchie's model: "Was that a joke?" (Typicality) and "How funny was it?" (Quality). JAPE's jokes were compared to human jokes and to two categories of human-generated non-jokes. Binsted *et al.* found that children rate human-generated jokes as more typical and of higher quality than non-jokes. JAPE's jokes were somewhere in between. Ritchie *et al.* performed further tests on this data and repeated the study with college students (Ritchie et al. 2008). Thir results were broadly the same, but there was low inter-rater reliability, especially on Quality.

While we focus on four specific metrics in our work, we do not mean to imply that these metrics represent four completely separate schools of thought. Instead, all four are influenced by each other and by prior work such as Boden's. What they all have in common is the idea of decomposing creativity into sub-concepts, then measuring creativity by somehow measuring and combining other criteria. For example, under Ritchie's model, if one can calculate the Typicality and Quality of a creative work, one can then (by some means, perhaps a complex one) calculate the work's level of creativity. This contrasts to the Consensual Assessment Technique, in which judges rate creativity however they see fit. The advantage of a metrical perspective is that it invites standardized quantitative calculation and avoids circularity. We use four metrics from the literature to represent a range of influential perspectives within the paradigm of metrical assessment. Our aim is to add to our understanding of metrical assessment of creativity as a whole.

## Experiment I

### Method

We tested 4 common metrics for creativity evaluation: Ritchie's model, Pease *et al.*'s novelty and value criteria, Colton's creative tripod, and the IDEA model. These metrics are easy to test on human poetry since they focus on the creative product and not on the process. Since none of these metrics have been put into a standardized questionnaire form, we constructed our own five-point Likert scale-based rating system for each. Each participant was only shown the questions for one of the four metrics, not all four. The questions we used are as follows:

**Ritchie's model**

- This resembles other poems I have read. *(Typicality)*
- This is a high quality poem. *(Quality)*
- I don't think this is a very good poem. *(Quality, reverse coded)*
- This is not a poem. *(Typicality, reverse coded)*

**Pease's criteria**

- This is a high quality poem. *(Value)*
- This poem is not like other poems I have seen before. *(Novelty)*
- I don't think this is a very good poem. *(Value, reverse coded)*
- This poem is clichéd. *(Novelty, reverse coded)*

**Colton's Creative Tripod**

- The author of this poem seems to have no trouble writing poetry. *(Skill)*
- The author of this poem is imaginative. *(Imagination)*
- The author of this poem understands how poetry works. *(Appreciation)*
- The author of this poem isn't very good at writing poetry. *(Skill, reverse coded)*
- The author of this poem isn't bringing anything new or different into the poem. *(Imagination, reverse coded)*
- The author of this poem doesn't really know anything about poetry. *(Appreciation, reverse coded)*

**IDEA model**

- I like this poem. *(Wellbeing)*
- I am willing to spend time trying to understand this poem. *(Cognitive Effort)*
- This poem makes me unhappy. *(Wellbeing, reverse coded)*
- This poem is not worth bothering with. *(Cognitive Effort, reverse coded)*

It should be noted that the construction of questions to represent abstract concepts from existing models is a potential source of error. For example, the IDEA model's Wellbeing criterion is based on like or dislike of a poem; it is not clear how an ideal reader would respond if they appreciated a poem but found it very sad. Appreciation in Colton's tripod, despite the lack of strict definitions of Colton's terms, also arguably refers to a creator's ability to evaluate its own work, rther than its ability to understand its field in general. However, researchers such as Norton et al (Norton, Heath, and Ventura 2010) refer to the Appreciation part of the tripod when training computers to apply labels to pre-existing images, implicitly lending support for the latter interpretation. After all, to evaluate one's own art one needs to be able to understand and evaluate art in general. A fully robust set of questions for a standardized questionnaire would require repeated testing and refinement in a variety of contexts; we have not yet reached the point of performing such tests.

### Data

For this experiment we used three hand-collected data sets of contemporary poetry written by humans. Each set contained 30 short poems in English of between 5 and 20 lines; we stuck to contemporary poetry so as to avoid different eras of poetry becoming a confounding factor, and so as to minimize the probability that a study participant had read the poems before. In no case did more than one poem by a single author appear across data sets.

For our purposes, we assumed that poems published in professional venues are more creative than poems written by novices. That is, we assumed that the editors of poetry magazines are experts and that their opinion strongly correlates with the actual creativity of the poetry published. This is, of course, debatable. Editors are sure to have specific cultural tastes and biases, but since all human judgments of creativity are culturally situated we find it an acceptable simplifying assumption.

The **Good** data set was composed of poems from Poetry Magazine between November 2013 and April 2014. Poetry Magazine is a very long-established, professional magazine which can reasonably be considered to contain the work of the most critically acclaimed mainstream literary poets working today. All poems meeting the length and non-duplication requirements and appearing in the magazine during this time window were selected, with the exception of a

| **Metric** | Criterion | Good | Medium | Bad | $F$ |
|---|---|---|---|---|---|
| Ritchie | Typicality | 0.20 | 0.41 | 1.23 | 10.6** |
| | Quality | 0.23 | 0.67 | 1.40 | 10.2** |
| IDEA | Wellbeing | 0.78 | 1.14 | 1.54 | 13.9** |
| | Cognitive Effort | 0.60 | 0.94 | 1.46 | 14.9** |
| Colton | Imagination | 0.75 | 1.16 | 1.07 | 2.3 |
| | Appreciation | 0.67 | 1.11 | 1.68 | 8.3** |
| | Skill | 0.44 | 0.84 | 1.40 | 7.4* |
| Pease | Novelty | 0.96 | 0.80 | 0.49 | 9.8** |
| | Value | 0.17 | 0.44 | 0.72 | 3.6 |

Table 1: Average ratings and $F$ scores for poem categories according to each metric. Each component is scored between -4 and +4. Significant results ($p < 0.05$) following Bonferroni correction are marked with a *, or ** if highly significant ($p < 0.01$).

few which were discarded due to complex visual formatting and two which were discarded due to experimenter discomfort. The remaining 30 poems comprised the Good data set.

The **Medium** data set was composed of 2 poems each from 15 lesser-known online magazines. Some of these were magazines devoted exclusively to poetry while others were a combination of poetry and prose. Each magazine pays a token amount (between US $5 and $10) per a poem. For each magazine, the most recent 2 poems meeting length and non-duplication requirements were chosen for the data set, with a single exception in which one poem was discarded and the third-most-recent poem chosen as a replacement. This added up to a Medium data set of 30 poems.

The **Bad** data set was composed of poems by unskilled amateur poets. We chose these poems by going to the Newbie Stretching Room at the Poetry Free-For-All, an online poetry critique forum. This section is for newcomers who have not posted poetry on the forum before; both experienced moderators and other newcomers can comment on the poems. We chose poems meeting the length and non-duplication requirements from this section, and discarded any which had received positive feedback from a moderator. Most of the chosen poems received comments from moderators instructing the author to read introductory articles on how to improve; a few had more specific, pointed comments. (Example: "This is dreadfully bad beginner's doggerel that fails for many, many, many reasons.") Selecting the most recently posted poems which fit these requirements resulted in a Bad data set of 30 poems.

Finally, we collected a **Test** data set containing 6 texts which were the same length as the chosen poems, but were obviously not poems. 3 of these were snippets from business news, and 3 from sports news.

These data sets are all available upon request.

## Collection

We recruited study participants on Crowdflower, a crowd-sourced microtasking website. In order to minimize cultural and linguistic difference as a confounding factor, participants were limited to those living in the United States.

Each participant was given six poems at a time, selected from any or all of the data sets, and shown the questions for only one of the four metrics. The participant was then asked to rate each poem based on that metric. Participants could rate poems repeatedly up to a maximum of 36 poems per participant per metric. We collected enough responses to amass 20 responses on each metric for each poem.

Participants were not shown the headings or names for the metric they were given, nor the names of the criteria on which the questionnaire items were based. Our justification for separating the metrics in this manner, and for coding Quality and Value separately even though the questions are identical, is that we were interested in taking each metrical approach as a whole, rather than mixing and matching criteria from all the metrics.

For each criterion, we ran a single-factor ANOVA comparing the Good, Medium, and Bad poems' scores on that criterion. Since there were three two-criterion metrics and one three-criterion metric, we ran nine ANOVAs and then applied a Bonferroni correction for nine hypotheses. The null hypothesis was that, for all metrics, participants' responses to Good, Bad, and Medium poems would be drawn from an identical distribution. The alternative hypothesis was that the distributions would not be identical: that is, that on some criteria, poems from one or more categories would be rated differently than others.

## Results

Results were the opposite of what we expected. For most criteria, participants rated Bad poems significantly (at $p = 0.05$ or better, following Bonferroni correction) more highly than Good ones. The exception was Novelty, in which Good poems were rated more highly than Bad. For Imagination and Value, the differences between categories were not significant. Exact F and p-values for each of these criteria are shown in Table 1.

This was a highly surprising result since it is not attributable to rater incompetence or failure to pay attention. Incompetent crowd workers who failed to pay attention might give the same score to all poems, or give random scores. Our raters, however, had significantly different reactions to the different groups. Adding test questions and bonuses to incite workers to pay more attention did not change the overall response pattern. This indicates that crowd workers can differentiate between these groups—but their preferences are different from what we had imagined.

The results for Medium poems were more ambiguous. We ran a Fisher's Least-Significant Difference Test to under-

| **Metric** | Criterion | C-Bad | C-Good | *t* |
|---|---|---|---|---|
| Ritchie | Typicality | 1.25 | 1.46 | 0.48 |
| | Quality | 0.08 | 0.77 | 0.13 |
| IDEA | Wellbeing | 1.30 | 1.61 | 0.35 |
| | Cognitive Effort | 0.11 | 0.63 | 0.23 |
| Colton | Imagination | 0.65 | 0.80 | 0.74 |
| | Appreciation | 1.12 | 1.50 | 0.42 |
| | Skill | 0.84 | 1.20 | 0.40 |
| Pease | Novelty | 0.32 | 0.24 | 0.74 |
| | Value | 0.11 | 0.34 | 0.62 |

Table 2: Average ratings and *t* scores for children's poem categories according to each metric. At $p < 0.05$, there were no significant differences found after Bonferroni correction

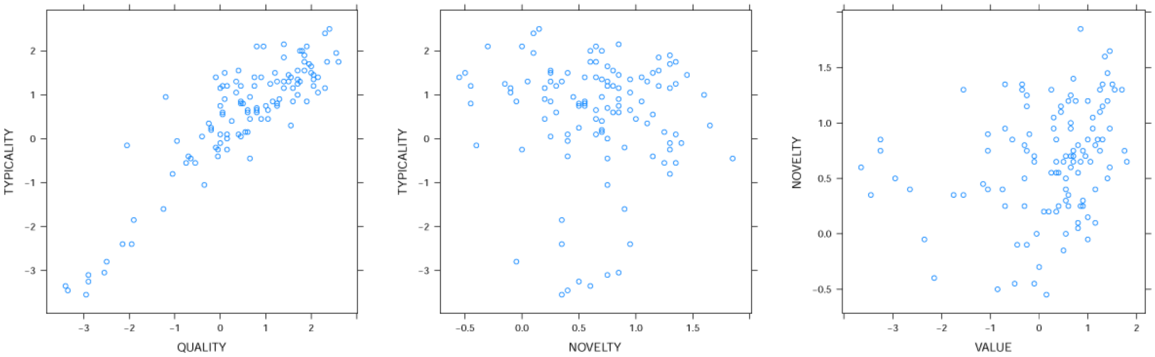

Figure 1: Sample scatterplots showing relationships between Novelty, Typicality, and Quality for poems in all of the data sets from both experiments.

stand the pairwise relationships between the three groups, again applying Bonferroni correction. Although Medium poems generally rated more highly than Good poems, in no case was this statistically significant. The difference between Medium and Bad poems, meanwhile, depends on the criterion. For Typicality, Novelty, and Effort, Medium poems were significantly different from Bad ones. For the other criteria, there was no significant difference between Medium poems and either other group.

## Experiment II

One potential explanation for why participants preferred Bad poems is that the Bad poems were more accessible. Poems from a prestigious literary journal may be difficult to understand due to heavy allusiveness and other poetic conventions. To test the inaccessibility hypothesis, we ran a second experiment focusing on poems written with children as the indended audience.

The **C-Good** data set was composed of children's poems found in the Children's Poetry section of the Poetry Foundation website in November 2014. The same selection constraints were used as with the first data set: poems were between 5 and 20 lines in length and no poet's work was used more than once. We also excluded poems by poets born prior to the 20th century. We collected a total of 10 C-Good poems, by authors such as Kenn Nesbitt and Shel Silverstein.

The **C-Bad** data set was composed of poems posted on the Family Friend Poems forum by amateur poets between September and November 2014, meeting the length and author uniqueness criteria. 10 such poems were selected. As there is no expectation of detailed critique at Family Friend Poems, we did not filter poems by critiques given as we did with the Bad adult poems. In fact, most responses to these poems were brief and complimentary (e.g. "Brilliant. Loved it 10"), even when the poems made large mistakes with meter and rhyme.

These poems were randomized and evaluated in the same way as the poems from Experiment I, on the criteria from the same four metrics. Since there are only two data sets in Experiment II, a *t*-test was performed on every criterion to detect differences in how the children's poems were rated.

### Results

The children's poem results lacked the effect seen in the adult poems. Participants rated C-Good poems more highly than C-Bad poems on most criteria, but these results were not statistically significant. A power analysis determined that this was not solely a result of the smaller size of the second study; hundreds of poems would have been needed for significance. Using children's poems removed raters' preference for bad poems, but did not introduce a preference for good poems above the level of noise.

## Correlations within and between metrics

It is not empirically clear if the different criteria from the different metrics actually elucidate different components of creativity. We investigated this by combining the data from Experiments 1 and 2, then generating scatterplots and correlation coefficients to examine the relationships between different criteria. With the exception of Novelty, all criteria were fairly well-correlated with each other ($0.65 < r < 0.99$), and scatterplots showed approximately linear relationships. Novelty had no significant positive, negative, or non-linear relationship with any other criterion. Example scatterplots are given in Figure 1

The high correlations between different criteria may indicate that these criteria—especially those with extremely high correlations, such as Skill and Appreciation at $r = 0.99$—are not actually separate concepts, or at least, are not adequately separated in the minds of raters when phrased as our questionnaire phrases them. An alternative interpretation, suggested by a reviewer to this paper, is that the high correlation is a good thing: if all criteria measure some aspect of creativity, then one would expect them all to change in similar ways along with an underlying change in creativity.

# Discussion

Our goal was to illustrate differences in effectiveness between different metrics, but we ended up finding something different. When using metrics, rather than simply asking judges how creative something is, the purpose is to be more objective and ensure that the appropriate factors are considered. However, the criteria we tested were not objective enough to produce trustworthy judgments from non-expert raters. Regardless of the criteria, non-expert raters showed a strong bias against Good poems due to these poems' inaccessibility. Even when more accessible poems were used, non-expert raters were unable to clearly distinguish between skillful and unskillful human poems.

## On Novelty, Typicality, Quality and Value

A major difference between Ritchie's (Ritchie 2001) and Pease *et al.*'s (Pease, Winterstein, and Colton 2001) work is the concept of Novelty. While Pease *et al.* define Novelty as a necessary component of creativity, Ritchie prefers to measure its opposite, Typicality. The claim is that, first, a creative computational system must learn to produce acceptable examples of the target output class. For example, a poetry program should not simply produce random words, but should produce something recognizeable as a poem. Only when this hurdle has been crossed can we begin to work towards novel forms of poem.

It is commonly claimed that the novelty and quality of creative works should form a Wundt curve. A completely non-novel work is not interesting. As works begin to diverge meaningfully from other works in their target class, they become more interesting. However, works which are too novel can be off-putting or difficult to accept. At the extreme, a completely novel and chaotic work is indistinguishable from meaningless noise, and is uninteresting for that reason. Therefore, an optimal creative work should involve a moderate amount of novelty. The empirical evidence for such a Wundt curve is not strong (see (Galanter 2012)) but when Ritchie and others treat typicality as a prerequisite to novelty, they implicitly argue for such a curve.

Our research fails to show a Wundt curve or similar relationship between novelty, typicality, and value. Indeed, our research suggests that typicality and novelty are not opposites: the correlation between them is nearly zero ($R = -0.05$). Poems with high Typicality may have high or low Novelty, and vice versa. Typicality is strongly correlated with most of the other criteria tested, with our non-expert raters seeing poems as more valuable, skillful, etc the more typical they are. Even though our data set included very atypical works (non-poems), there did not appear to be a threshold at which poems became "typical enough" for novelty to become relevant.

Meanwhile, Good poems are rated as more novel than Bad. Taken at face value, this would suggest that Novelty might be a better metric than others for measuring creativity. However, the effect for novelty disappears when applied to children's poems. Rather than measuring the creativity of a poem, it is more likely that Novelty for non-expert raters measures inaccessibility: Good poems are rated as more novel than others because they are more difficult to understand. This implies that a participant's rating of a poem as novel may signify discomfort. Without enough domain expertise to see the meaning underlying novelty, non-expert judges prefer poems without it.

## On accessibility and the target audience

If non-expert judges prefer a minimum of novelty, one would expect to see a very different pattern of response from experts. If a poem can be too novel, then this raises the question: too novel *to whom?* Clearly, to the editors of Poetry Magazine, each poem in their magazine made sense and was of high quality. Yet Crowdflower users—presumably ordinary people with little formal education in poetry—saw less quality and sense in these poems than in the work of novice poets.

The poems in Poetry Magazine are so complex that the magazine comes with an explanatory Discussion Guide. Poems allude heavily to other works and imply or illustrate things instead of stating them outright; some raise difficult questions such as "who is creating what, as well as who is inside the work and who is outside" (Poetry Foundation 2014). Without education in poetry, it is no wonder that an ordinary person finds such complexity offputting. Our results suggest that this offputting effect may be so strong that it drowns out any other differences between skilled and unskilled human poetry. To non-expert judges, the confusing complexity of professional poems is worse than any of the clumsiness of an amateur. Yet to an expert in poetry, it would be absurd to say that the amateur poems are therefore of higher quality.

The strength of the effect here—not just negating but reversing expected trends—is surprising. It suggests that there is a great danger in ignoring the question of rater expertise. The use of specific criteria such as Novelty, Value, Skill, Appreciation, or Imagination does not remove the need for this question. When poems are judged for their quality, who performs that judgment? The researcher? An ordinary reader? An expert? If so, what kind of expert? Future computational creativity studies need to make their answers to these questions explicit, even if they are not already using techniques which demand the use of experts.

In the meantime, without an identifiable target audience, it may be very dangerous to talk about quality, value, or skill in computational creativity as though it is only one thing. The quality of popular appeal and the quality of appeal to experts may be diametrically opposed, and there may be other audiences with still other views of quality. Until such an audience is chosen and the choice justified, the notion of creativity, without the notion of creativity *to whom*, is operationally meaningless.

## Conclusions

Using the conceptual criteria from four popular computational creativity evaluation metrics, we have shown that non-expert humans using these metrics can produce the opposite result from what is intended. Non-expert humans prefer more accessible poetry, even if that poetry is much less skilled according to experts. These results strongly suggest that even when structured metrics are being used, non-expert judges cannot approprately evaluate the creativity of a human or computer system. Regardless of the metric used, care must be taken in selecting and assessing an appropriate group of judges.

## References

Aguilar, W., and Pérez y Pérez, R. 2014. Criteria for evaluating early creative behavior in computational agents. In *Proceedings of the Fifth International Conference on Computational Creativity*, 284–287.

Amabile, T. 1983. In *The social psychology of creativity*, 37–64. Springer-Verlag New York.

Binsted, K.; Pain, H.; and Ritchie, G. 1997. Children's evaluation of computer-generated punning riddles. *Pragmatics & Cognition* 5(2):305–354.

Boden, M. A. 1990. *The creative mind: Myths and mechanisms*. Psychology Press.

Bown, O. 2014. Empirically grounding the evaluation of creative systems: incorporating interaction design. In *Proceedings of the Fifth International Conference on Computational Creativity*.

Brown, D. 2009. Computational artistic creativity and its evaluation. Number Dagstuhl Seminar 09291, 1–8.

Burns, K. 2006. Atoms of eve: A bayesian basis for esthetic analysis of style in sketching. *AIE EDAM: Artificial Intelligence for Engineering Design, Analysis, and Manufacturing* 20(03):185–199.

Burns, K. 2012. Eve s energy in aesthetic experience: a bayesian basis for haiku humour. *Journal of Mathematics and the Arts* 6(2-3):77–87.

Burns, K. 2015. Computing the creativeness of amusing advertisements: A Bayesian model of Burma-Shave's muse. *AIE EDAM: Artificial Intelligence for Engineering Design, Analysis, and Manufacturing* 29(01):109–128.

Chan, H., and Ventura, D. 2008. Automatic composition of themed mood pieces. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, 109–115.

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

Colton, S. 2008b. Experiments in constraint-based automated scene generation. *Proceedings of the Fifth International Workshop on Computational Creativity 2008* 127.

Das, A., and Gambäck, B. 2014. Poetic machine: Computational creativity for automatic poetry generation in bengali. In *Proceedings of the Fifth International Conference on Computational Creativity*.

Galanter, P. 2012. Computational aesthetic evaluation: Past and future. In *Computers and Creativity*. Springer. 255–293.

Gervás, P. 2002. Exploring quantitative evaluations of the creativity of automatic poets. In *Proc. of the 2nd Workshop on Creative Systems, Approaches to Creativity in Artificial Intelligence and Cognitive Science, the 15th European Conf. on Artificial Intelligence (ECAI 2002)*.

Gervás, P. 2007. On the fly collaborative story-telling: Revising contributions to match a shared partial story line. *International Joint Workshop on Computational Creativity* 13.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Karampiperis, P.; Koukourikos, A.; and Koliopoulou, E. 2014. Towards machines for measuring creativity: The use of computational tools in storytelling activities. In *Proceedings of the 14th International Conference on Advanced Learning Technologies (ICALT)*, 508–512.

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.

Lehman, J., and Stanley, K. O. 2012. Beyond open-endedness: Quantifying impressiveness. In *Artificial Life*, volume 13, 75–82.

Llano, M. T.; Hepworth, R.; Colton, S.; Gow, J.; Charnley, J.; Granroth-Wilding, M.; and Clark, S. 2014. Baseline methods for automated fictional ideation. In *Proceedings of the Fifth International Conference on Computational Creativity*, 211–219.

Monteith, K.; Brown, B.; Ventura, D.; and Martinez, T. 2013. Automatic generation of music for inducing physiological response. In *Annual Meeting of the Cognitive Science Society*, 3098–3103.

Monteith, K.; Martinez, T.; and Ventura, D. 2010. Automatic generation of music for inducing emotive response. In *Proceedings of the International Conference on Computational Creativity*, 140–149.

Norton, D.; Heath, D.; and Ventura, D. 2010. Establishing appreciation in a creative system. In *Proceedings of the International Conference on Computational Creativity*, 26–35.

Norton, D.; Heath, D.; and Ventura, D. 2013. Finding creativity in an artificial artist. *The Journal of Creative Behavior* 47(2):106–124.

Pearce, M. T., and Wiggins, G. A. 2007. Evaluating cognitive models of musical composition. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, 73–80.

Pease, A.; Winterstein, D.; and Colton, S. 2001. Evaluating machine creativity. In *Workshop on Creative Systems, 4th International Conference on Case Based Reasoning*, 129–137.

Poetry Foundation. 2014. Poetry magazine discussion guide. http://www.poetryfoundation.org/poetrymagazine/guide/89. Accessed: 2015-02-03.

Rashel, F., and Manurung, R. 2014. Pemuisi: a constraint satisfaction-based generator of topical Indonesian poetry. In *Proceedings of the Fifth International Conference on Computational Creativity*, 82–90.

Riedl, M. O., and Young, R. M. 2006. Story planning as exploratory creativity: Techniques for expanding the narrative search space. *New Generation Computing* 24(3):303–323.

Ritchie, G.; Munro, R.; Pain, H.; and Binsted, K. 2008. Evaluating humorous properties of texts. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, 17.

Ritchie, G. 2001. Assessing creativity. In *Proc. of AISB01 Symposium*.

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

Román, I. G., and y Pérez, R. P. 2014. Social Mexica: A computer model for social norms in narratives. In *Proceedings of the Fifth International Conference on Computational Creativity*, 192–200.

Smith, M. R.; Hintze, R. S.; and Ventura, D. 2014. Nehovah: A neologism creator nomen ipsum. In *Proceedings of the Fifth International Conference on Computational Creativity*, 193–181.

Tearse, B.; Mawhorter, M. M. P.; and Wardrip-Fruin, N. 2011. Experimental results from a rational reconstruction of MINSTREL. In *Proceedings of the Second International Conference on Computational Creativity*.

Ventura, D. 2008. A reductio ad absurdum experiment in sufficiency for evaluating (computational) creative systems. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, 11–19.

Young, M. W.; Bown, O.; et al. 2010. Clap-along: A negotiation strategy for creative musical interaction with computational systems. In *Proceedings of the International Conference on Computational Creativity 2010*, 215–222.

## Poetry Sources

Bell, J., and Ius, D. Vine Leaves Literary Journal. http://www.vineleavesliteraryjournal.com/. [accessed April 2014].

Bobet, L. Ideomancer. http://www.ideomancer.com/. [accessed April 2014].

Card, O. S. Strong Verse. http://www.strongverse.org/. [accessed April 2014].

Delmater, W. S. Abyss & Apex. http://www.abyssapexzine.com/. [accessed April 2014].

el Mohtar, A., and Paxson, C. Goblin Fruit. http://www.goblinfruit.net. [accessed April 2014].

ELJ Publications. Amethyst Arsenic. http://www.amethystarsenic.com/. [accessed April 2014].

Gage, K., and Filek, M. K. Writing Tomorrow. http://writingtomorrow.com/. [accessed April 2014].

Gaskin, E. Astropoetica. http://www.astropoetica.com/. [accessed April 2014].

Greene, R. Raleigh Review. http://www.raleighreview.org/. [accessed April 2014].

Hart, M. Through the Gate. http://throughthegate.net/. [accessed April 2014].

Peg Leg Publishing. GlassFire Magazine. http://www.peglegpublishing.com/glassfire.htm. [accessed April 2014].

Poetry Foundation. Poetry Magazine. http://www.poetryfoundation.org/poetrymagazine/. [accessed April 2014].

Poetry Free-For-All, T. Newbie Stretching Room. http://www.everypoet.org/pffa/forumdisplay.php?26-Newbie-Stretching-Room. [accessed April 2014].

Rademacher, K. Silver Blade. http://silverblade.silverpen.org/. [accessed April 2014].

Unknown publisher. Neon - A Literary Magazine. http://neonmagazine.co.uk/?p=5103. [accessed April 2014].

Well Done Marketing, Inc. Punchnel's. http://www.punchnels.com/. [accessed April 2014].