Assessing the Creativity of Designs at Scale

Christopher J. MacLellan
HCII, Carnegie Mellon University
Pittsburgh, PA
cmaclell@cs.cmu.edu

ABSTRACT

How best to assess the creativity of a large number of designed artifacts remains an open problem. The typical approach is to have experts answer likert questions about individual artifacts. This process typically requires a substantial amount of training to ensure the judges achieve an acceptable level of agreement. Consequently, the approach does not scale well as it is infeasible to have multiple experts regularly evaluate the creativity of a large number of designs. The current work explores an alternative approach that uses both individual and pairwise judgements from novice crowd workers to support reliable and scalable assessment of creative designs. This approach, which we call TrueCreativity, can operate over a set of evaluations from a large number of judges and appropriately weights their evaluations based on their past reliability and agreement with other judges. We show that this approach produces results that strongly correlate with another measure of creativity.

Author Keywords

Empirical Methods, Quantitative; E-Learning and Education; Machine Learning and Data Mining; Crowdsourcing

ACM Classification Keywords

H.5.3 [Information Interface and Presentation]: Group and Organization Interfaces – Evaluation/methodology.

INTRODUCTION

The ability to assess the creativity of designed artifacts in a scalable way has implications for both design research and instruction. Scalable techniques for creativity assessment might be used to conduct larger-scale empirical studies into creativity, facilitating the process of scientific discovery. Further, techniques for the scalable assessment of artifact creativity have implications for how design is taught. For example, Scott Klemmer recently launched a massive open online course on Human-Computer Interaction. A key component of this course was getting grades on designed artifacts from peers [1]. Peer grading allowed students in the course to get feedback in situations where it would be infeasible for the instructors to grade every design. A scalable technique for measuring the creativity of designed artifacts might be used to facilitate this peer grading process.

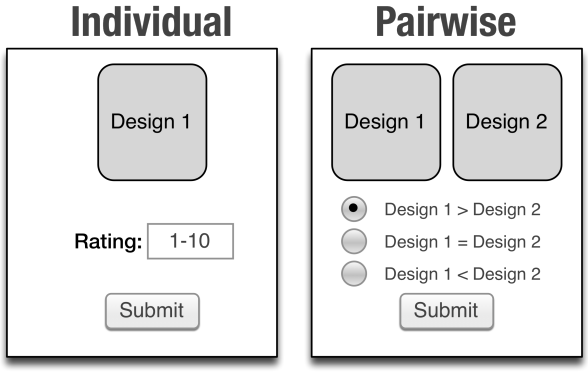

Figure 1. Simplified depictions of the rating formats used in the individual and pairwise rating schemes.

Despite the benefits, there is no clear approach to assessing the creativity of designs at scale. Traditionally, design creativity is assessed by a panel of experts using a likert measure. Scaling this approach consists of having novices, instead of experts, complete the same measures [1]. However, there are several criticisms against this approach [2]. In particular, each judge may award higher or lower ratings on average or they may award the same average rating, yet discriminate more finely amongst the designs. The typical approach to overcoming these difficulties is to train the judges until they achieve acceptable agreement, and to regularly retrain them to prevent rater drift. However, this is infeasible when using a large number of raters (e.g., mechanical turk workers or peers from an online class). Recent approaches have explored the use of reference items to correct for judge bias [5], but this approach still does not account for differences in judge discrimination abilities and it is unclear how the selection of reference items impacts the estimation of judge bias.

TRUECREATIVITY

To overcome these challenges we developed TrueCreativity, a Bayesian method for reliably assessing the creativity of designed artifacts using novices from the crowd. To apply this method we collect both individual creativity ratings (we used a 10 point scale) and pairwise creativity ratings (i.e., asking a judge to determine which of two designs is more creative or if they are equal) from crowd workers. Figure 1 depicts the rating formats we had judges use. These ratings are then combined into a single estimation of each design's latent creativity using a statistical model that estimates and corrects for judge bias and discrimination in both rating formats. We chose to support pairwise ratings because they do not suffer

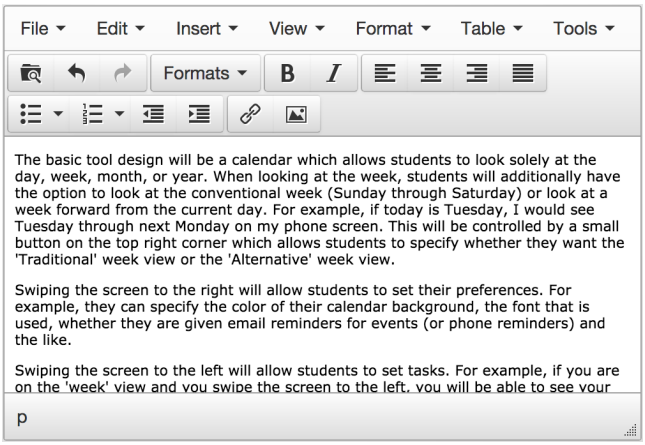

Figure 2. TinyMCE interface with an example participant design.

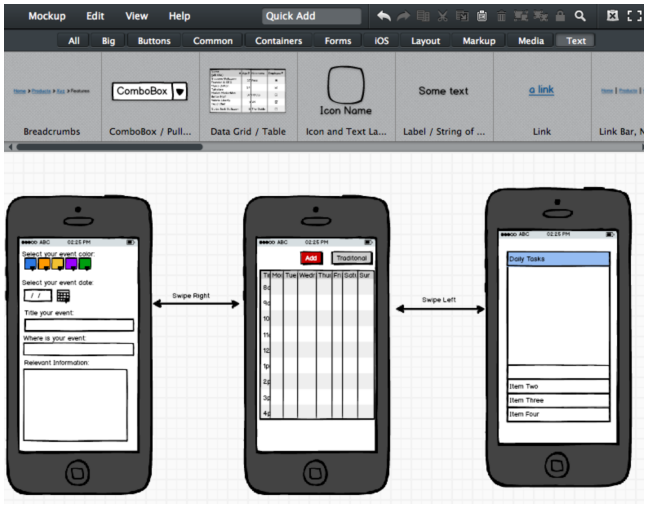

Figure 3. Balsamiq interface with an example participant design.

from the same criticisms as individual ratings (i.e., pairwise ratings are scale independent) and because research indicates that ordinal rating is easier and more reliable for novices [3]. To combine both formats of ratings we first specified the likelihood of each rating given the item and judge parameters. We modeled the likelihood of the individual ratings using a linear model that has a parameter for each item (i.e., TrueCreativity) and two parameters for each judge (i.e., bias and discrimination). For the pairwise ratings, we modeled the likelihood of each rating using a multi-class logistic regression that has one parameter for each item being compared (i.e., TrueCreativity) and three parameters for each judge (i.e., their lower and upper thresholds for rating items as equal and discrimination). After specifying the models, we used Markov Chain Monte Carlo optimization to compute the most likely TrueCreativity, bias, and discrimination values given all of the ratings in both formats. This method extends prior approaches that correct for judge bias [5]. In particular, we eliminate the need for reference items by using all of the ratings to jointly estimate judge parameters and item parameters. In essence, we use all overlapping ratings across both models to estimate and correct for judge bias and discrimination abilities.

PRELIMINARY RESULTS

As a preliminary evaluation of our approach we assessed the creativity of two sets of smart phone time management application designs. One set consisted of wireframes created using Balsamiq (http://www.balsamiq.com). The other set consisted of textual designs produced using a word processor. Figures 2 and 3 shows an example design in each format. For each design we had an existing creativity measure that was based on the number of unique features present in the designs. This "feature count" measure was produced by two expert coders who first generated a list of 18 unique features by studying the designs and then independently coded the designs in terms of the features they contained. For each feature the agreement of the coders was high (Cohen's $\kappa > 0.7$). This approach is similar to the ideation quantity measures used in other studies of creativity [4].

To compute the TrueCreativity scores we had 36 workers from Amazon Mechanical Turk and 6 researchers from our lab independently judge the creativity of the designs using both rating schemes. For each wireframe we collected 6 individual ratings and 18 pairwise comparisons. For each textual design we collected 3 individual ratings and 9 pairwise comparisons. After computing the TrueCreativity measure using these ratings, we found that it was strongly correlated with the feature count measure (Pearson's $\rho = 0.5, p < 0.01$ for the wireframes and Pearson's $\rho = 0.72, p < 0.01$ for the textual designs). This agreement suggests that the measures have good convergent validity. While it took approximately one month to develop and achieve a reliable coding scheme for the feature count measure, it only took two days to collect the ratings necessary for computing the TrueCreativity scores. Given these results, future work will focus on rigorously assessing the reliability and validity of this approach.

ACKNOWLEDGEMENTS

This work is advised by Ken Koedigner and Steven Dow and supported by IES (R305B090023) and NSF (IIS-1208382 and IIS-1217096).

REFERENCES

1. Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. R. Peer and Self Assessment in Massive Online Classes. *TOCHI 9*, 4 (2014), 131–168.
2. Pollitt, A. Comparative judgement for assessment. *International Journal of Technology and Design Education 22*, 2 (2012), 157–170.
3. Raman, K., and Joachims, T. Bayesian Ordinal Peer Grading. In *L@S '15* (2015), 149–156.
4. Shah, J. J., Kulkarni, S. V., and Vargas-Hernandez, N. Evaluation of Idea Generation Methods for Conceptual Design: Effectiveness Metrics and Design of Experiments. *JMD 122*, 4 (2000), 377–384.
5. Xu, A., and Bailey, B. P. A Reference-Based Scoring Model for Increasing the Findability of Promising Ideas in Innovation Pipelines. In *CSCW '12* (2012), 1183–1186.