Proceedings of ASME 2018 International Design Engineering Technical Conferences & Design Conference
August 18-21, 2019, Anaheim, California

DRAFT - IDETC2019 – 97710

CALIBRATING NOVELTY METRICS USING HUMAN RATERS: A CASE STUDY

Sharath Kumar Ramachandran[1]
[1] **School of Engineering Design, Technology and Professional Programs**
The Pennsylvania State University,
State College, Pennsylvania, 16802
Email: mqa5244@psu.edu

Faez Ahmed
Dept. of Mechanical Engineering
University of Maryland
College Park, Maryland, 20742
Email: faez00@umd.edu

Mark Fuge
Dept. of Mechanical Engineering
University of Maryland
College Park, Maryland, 20742
Email: fuge@umd.edu

Samuel T. Hunter
Dept. of Psychology
The Pennsylvania State University,
State College, Pennsylvania, 16802
Email: sth11@psu.edu

Scarlett R. Miller[1,2]
[2] **Dept. of Industrial and Manufacturing Engineering**
The Pennsylvania State University,
State College, Pennsylvania, 16802
Email: scarlettmiller@psu.edu

INTRODUCTION

One of the key challenges vexing engineering design researchers is assessing the inherently nebulous construct of creativity in design studies [1]. Amongst the various aspects of creativity, originality or novelty is most recognized. This is because research has shown that individuals exhibiting creativity are generally non-adhering, unorthodox, and autonomous when it comes to solving problems [2]. In an engineering context, research has shown that there is a higher possibility of solving a design problem when a more discrete or original ideas is produced in the initial stages of the design process [3]. Because of this, engineering researchers have long sought to identify what helps individuals generate novel concepts, which necessitates the need to measure design novelty.

In engineering, there have been two vastly different approaches adopted to measure concept novelty; subjective ratings based on human judges (see for example [4, 5]) and quantifiable methods based on feature-trees (see for example[6-17]). The use of subjective ratings in engineering design research stems from the early work of Psychologist Teresa Amabile [18-20] who developed the consensual assessment technique, which is based on the underlying premise that an idea is creative (novel, useful, meaningful) only to the extent to which raters *independently* agree that it is creative. In contrast to this approach, *quantifiable* measures in engineering design typically rely on breaking down design concepts into their components and then quantifying the novelty of each of these components. The 'gold standard' metric in this area was developed by Shah, Vargas-Hernandez, and Smith (SVS) [21] who proposed a feature tree based method that differentiates novelty based on the physical principle, working principle, embodiment, and detail levels and assigns a relative importance (or weight) of 10, 6, 3, and 1 to these categories. However, research has shown that applying these two different approaches to the same design problem can result in creativity rankings that are not only vastly different, but often negatively correlated [22, 23]. The question is, why?

The goal of this work is not to identify which metric was 'superior' to the other. Instead, it is meant to provide a method for leveraging human raters as a means to calibrate the more automated-feature tree approaches. As a first step towards this goal, it is important to decipher what characteristics designers are using to distinguish design concepts and how this differs from or is similar to more feature-tree based approaches.

METHODOLOGY

In order to achieve this goal, a study was conducted where we asked four raters to create an idea map based on the similarity

of 10 ideas using a think-out-loud protocol, see Figure X for example map. Specifically, each participant was provided with the same 10 idea sketches, printed on 8.5" x 5.5" sheets of paper. The order of the ideas was randomized for each participant. The participants were asked to pin the sketches on a 65" x 55" canvas, such that the distance between any two sketches would be proportional to how similar the ideas were to each other. The sketches were allowed to overlap. The subjects were allowed to move the sketches multiple times, until they were satisfied with the idea map created. The participants were allotted a maximum time of 30 minutes for the activity. Participants were required to think aloud as they placed and moved the ideas around on the canvas. Throughout the activity, the participants were recorded using audio and video equipment. Transcription of the audio files were completed using NVivo transcription services and the automatic transcriptions were manually corrected for accuracy. Once the transcriptions were complete, NVivo Pro version 12.0 was used to analyze the data adapting principles of inductive content analysis. Specifically, the audio transcriptions were first reviewed and then the coders classified each sentence from the transcription into an open code using principles of inductive content analysis. Once open coding was completed, the categories were then grouped under higher order headings by collapsing those that were similar.

RESULTS

The inductive content analysis resulted in three higher-level codes; method of frothing (f = 91), form used (f = 83), and power source used (f = 34). For method of frothing, the top categories participants discussed were: *air* (f=28), *spinning* (f=33), and *vibrating* (f=11). For example, participant 1 stated, *"it has some an air intake and idea number two has air. So, I would say that that's probably pretty close. It's got something in common."* This was categorized as ideas being differentiated based on the method of frothing: air. On the other hand, the top codes that contributed to form included: *bicycle (f=17), cup* (f=16), *beaters* (f=14). For example, one participant said "It's pretty similar to idea number three because they both have these pedals and they connect to a frother. One is a bike, one is a pedal but they're pretty similar." This was categorized as being differentiated based on the form: bike, pedal. Finally, the top codes that contributed to power source category included: *manual* (f=22), *electric* (f=11). For example, one participant said, "some have either like a human or a bike powering them and then others just kind of seem to go on their own."

SVS levels	SVS weights	Content Analysis Theme	Weights from content analysis (10 scale)
Physical Principle	10	Method of frothing	10
Working Principle	6	Power source used	3.74
Embodiment	3	Form	9.12

Table 1. Comparison of weights obtained from content analysis and weights provided by the SVS novelty metric.

The frequency of these codes were then normalized to identify the relative importance of these categories for the human raters, see Table 1. As can be seen, the method of frothing was the most frequently mentioned category followed by the form of the frother and then the power source used. When compared to the weights used in SVS [21], we see a clear discrepancy in the assignment of the prevalence or weight of these categories. Our prior work has identified a negative correlation between human ratings and the SVS feature-tree method. This difference in the relative importance of these categories may be contributing to the large differences between feature-tree methods and human raters. However, we may be able to utilize the method described here to modify feature-tree based methods using the relative importance gathered from human raters. For example, human raters can be used to identify categories that are important in novelty ratings, and the relative importance of these categories. This data can then be used to scientifically justify how feature-tree are created as well as provide justification for the weights.

CONCLUSIONS AND FUTURE WORK

The study presented here provides a methodology for using human raters to extrapolate the important categories and the relative importance of these categories for use in feature-tree based novelty methods. While the methodology presented here is novel, the utility of this approach needs to be verified in future work. The authors are currently in the process of comparing feature-tree based methods both before- and after- the implementation of the methodology above to identify if this method improves the alignment of human and feature-tree based methods. In addition, future work will be geared at verifying this approach in a variety of design contexts.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1728086.

REFERENCES

[1] Liikkanen, L. A., Hämäläinen, M. M., Häggman, A., Björklund, T., and Koskinen, M. P., "Quantitative evaluation of the effectiveness of idea generation in the wild," Proc. International Conference on Human Centered Design, Springer, pp. 120-129.

[2] Barron, F., and Harrington, D. M., 1981, "Creativity, intelligence, and personality," Annual review of psychology, 32(1), pp. 439-476.

[3] Henderson, D., Helm, K., Jablokow, K., McKilligan, S., Daly, S., and Silk, E., "A Comparison of Variety Metrics in Engineering Design," Proc. ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers.

[4] D'souza, N., and Dastmalchi, M. R., 2016, "Creativity on the move: Exploring little-c (p) and big-C (p) creative events within a multidisciplinary design team process," Design Studies, 46, pp. 6-37.

[5] Nikander, J. B., Liikkanen, L. A., and Laakso, M., 2014, "The preference effect in design concept evaluation," Design Studies, 35(5), pp. 473-499.

[6] Shah, J. J., Smith, S. M., and Vargas-Hernandez, N., 2003, "Metrics for measuring ideation effectiveness," Design studies, 24(2), pp. 111-134.

[7] Wilson, J. O., Rosen, D., Nelson, B. A., and Yen, J., 2010, "The effects of biological examples in idea generation," Design Studies, 31(2), pp. 169-186.

[8] Nelson, B. A., Wilson, J. O., Rosen, D., and Yen, J., 2009, "Refined metrics for measuring ideation effectiveness," Design Studies, 30(6), pp. 737-743.

[9] Verhaegen, P.-A., Vandevenne, D., Peeters, J., and Duflou, J. R., 2013, "Refinements to the variety metric for idea evaluation," Design Studies, 34(2), pp. 243-263.

[10] Verhaegen, P.-A., Peeters, J., Vandevenne, D., Dewulf, S., and Duflou, J. R., 2011, "Effectiveness of the PAnDA ideation tool," Procedia Engineering, 9, pp. 63-76.

[11] Duflou, J., and Verhaegen, P.-A., 2011, "Systematic innovation through patent based product aspect analysis," CIRP annals, 60(1), pp. 203-206.

[12] Schmidt, L. C., Hernandez, N. V., and Ruocco, A. L., 2012, "Research on encouraging sketching in engineering design," AI EDAM, 26(3), pp. 303-315.

[13] Doboli, A., and Umbarkar, A., 2014, "The role of precedents in increasing creativity during iterative design of electronic embedded systems," Design Studies, 35(3), pp. 298-326.

[14] Wodehouse, A., and Ion, W., 2012, "Augmenting the 6-3-5 method with design information," Research in Engineering Design, 23(1), pp. 5-15.

[15] Jagtap, S., Larsson, A., Hiort, V., Olander, E., and Warell, A., 2015, "Interdependency between average novelty, individual average novelty, and variety," International Journal of Design Creativity and Innovation, 3(1), pp. 43-60.

[16] Verhaegen, P.-A., Vandevenne, D., Peeters, J., and Duflou, J. R., 2015, "A variety metric accounting for unbalanced idea space distributions," Procedia engineering, 131, pp. 175-183.

[17] Coelho, D. A., and Vieira, F. L., 2018, "The effect of previous group interaction on individual ideation novelty and variety," International Journal of Design Creativity and Innovation, 6(1-2), pp. 80-92.

[18] Amabile, T., 1982, "Social psychology of creativity: A consensusual assessment technique," Journal of Personality and Social Psychology, 43, pp. 997-1013.

[19] Amabile, T., 1983, "Brilliant but cruel: perceptions of negative evaluators," Journal of Experimental Psychology, 19(2), pp. 146-156.

[20] Amabile, T., 1996, Creativitiy in Context, Westview Press, Boulder, Colorado.

[21] Shah, J., Smith, S., and Vargas-Hernandez, N., 2003, "Metrics for Measuring Ideation Effectiveness," Design Studies, 24, pp. 111-124.

[22] Gosnell, C. A., and Miller, S. R., 2016, "But is it creative? Delineating the Impact of Expertise and Concept Ratings on Creative Concept Selection.," Journal of Mechanical Design, 138(2)(2), pp. 021101-021101- 021101-021111.

[23] Sarkar, P., and Chakrabarti, A., 2011, "Assessing design creativity," Design Studies, 32(4), pp. 348-383.