Let’s not get too creative!

David C. Brown
Worcester Polytechnic Institute, Worcester, MA 01609, USA

This short position paper urges researchers to assume that existing research on creativity evaluation should form the basis for establishing computable metrics: what already exists, even if it is not obviously computable, should be examined carefully.

1. Introduction

The goal of this workshop is to encourage investigation of the different methodologies, theories and models currently employed to evaluate creativity, with a focus on those that are computable. This short position paper considers only metrics that can be used to predict that a designed *product* will be evaluated as being creative by humans.

Given the current published evidence we assume that *multiple* metrics need to be considered in order to adequately include all factors that influence the judgment of creativity by humans. We also assume that existing research on creativity evaluation should form the basis for establishing computable metrics: i.e., let’s not get too creative!

We start by reminding ourselves of some questions:

- What creativity metrics have been proposed?
- How have the metrics been evaluated?
- How accurately do individual metrics need to predict the ingredients of creativity to be useful?
- How have different metrics been combined?
- How accurately do combined metrics need to predict creativity to be useful?
- Which metrics have been converted from human use to computational use?

Design Computing and Cognition DCC’14. J.S. Gero (ed),
pp. xx-yy. © Springer 2014

Finding answered to these questions will take a lot of research, but we argue that this research is necessary. If psychologists, marketing theorists, design theorists and others have done serious research into creativity metrics, then we should respect that work, investigate its utility, and build on it.

Having a diverse set of measures allows use of an appropriate measure for a specific circumstance: e.g., a quickly calculated metric as a guide when aiming for a creative product, or a specific metric that works with a particular class of product or potential buyer. Some metrics may work well with heavily engineering-based products, while others work better with more decorative and stylish artifacts. It is clear that metrics only make sense in context [Brown 2014a].

As an example, in an informal experiment with graduate students in WPI's CS540 "AI in Design" course, students used their own invented metrics, as well as Besemer's [2006], to evaluate a variety of self-chosen artifacts including electrical tape, an ice scraper, and a Prius car. Depending on the artifact they reported having different kinds of difficulties.

2. Some existing Metrics

In this short paper we will merely *sample* some of the existing metrics, with an emphasis on diversity. The hope is that other researchers will take this challenge to complete the list (if that's possible), add more attributes, answer the questions given above, and, most importantly, determine each metric's utility for Computational Design Creativity (CDC) systems.

2.1. Besemer

Besemer [2006] has been researching and experimenting with creativity metrics from about 1980 `<http://ideafusion.biz/home/about-2/bibliography>`. She has used the Creative Product Semantic Scale (CPSS) tool over the last decade to evaluate the creativity of products as part of her consulting business. CPSS is based on the Creative Product Analysis Model (CPAM). It includes the three dimensions that have been found to be the "most important indicators of creativity in products": **Novelty**, **Resolution**, and **Style**. Novelty is broken into facets Surprising and Original; Resolution is broken into facets Logical, Useful, Valuable, and Understandable; and Style is broken into facets Organic, Well-crafted, and Elegant. While the CPSS tool does produce numerical output for all facets, it is intended for human use, and any equations used to map CPSS answers to facet scores are proprietary. See Brown [2013] for further discussion.

2.2. Cropley & Kaufman

Cropley & Kaufman [2012] propose the Creative Solution Diagnosis Scale (CSDS) with 30 indicators of creativity that they experimentally reduced to a Revised CSDS (RCSDS) with 24 indicators. This is a very subtle set of indicators that isolate some issues that most other developers of metrics gloss over. While additional metrics allow detection of subtle differences they may also make judgments more difficult.

Their five categories of indicators include: **Relevance/Effectiveness**, **Problematization**, **Propulsion**, **Elegance**, and **Genesis**. Relevance/Effectiveness is broken into indicators Performance, Appropriateness, and Correctness. Problematization is broken into indicators Prescription, Prognosis, and Diagnosis. Propulsion is broken into indicators Redefinition, Reinitiation, Generation, Redirection, and Combination. Elegance is broken into indicators Pleasingness, Completeness, Sustainability, Gracefulness, Convincingness, Harmoniousness, and Safety. Genesis is broken into indicators Vision, Transferability, Seminality, Pathfinding, Germinality, and Foundationality.

The RCSDS model reduces the categories to **Relevance/Effectiveness** (Performance, Appropriateness, and Correctness), **Novelty (**Problematization, Existing Knowledge, and New Knowledge**), Elegance (**Internal, External**)** and **Genesis** (Vision, Transferability, Seminality, Pathfinding, Germinality, and Foundationality).

Notice that Relevance/Effectiveness covers functionality and utility, much like Besemer’s Resolution category. The RCSDS also covers internal and external elegance (i.e., Style). It is odd not to see Novelty/Originality and Surprise on the list. However, these are partially covered under Problematization which includes how a product points out problems with, and improvements to, what exists. It does downplay “surprise” however, which is unfortunate.

What is distinctly different is Propulsion, which is about the use of new perspectives and new approaches, with a hint that “novelty” is due to those new directions. Even more different is Genesis, which is about how the novelty can indicate and influence future creativity. Note the impact of time and knowledge on this judgment, as looking back at the actual influence is quite different from anticipating it, where the level of expertize plays a large part in accuracy, but doesn’t guarantee correctness.

2.3. Dean et al.

Dean et al. [2006] examine a lot of non-design literature on creativity, including early work by Besemer. After refinement and testing they arrive at the dimensions **Novelty**, **Workability**, **Relevance**, and **Specificity**. Novelty is broken into the factors Originality and Paradigm Relatedness. Workability is broken into the factors Acceptability and Implementability. Relevance is broken into the factors Applicability and Effectiveness. Specificity is broken into the factors Implicational Explicitness, Completeness and Clarity (although after statistical testing Clarity was dropped from the system).

Dean et al. include rarity and being surprising under Novelty, as well as “the degree to which an idea preserves or modifies a paradigm”. They consider a new concept/artifact to consist of elements and relationships (including those with users). “Refining” a paradigm keeps both elements and relationships, while “Extending” comes from using new elements. Changing the relationships with the same elements results in “Redesign”, while one “Transforms” a paradigm by changing both. Note the connection to definitions of types of designing, as well as to Exploratory and Transformational creativity.

Workability is concerned with whether the design violates too many known constraints to be implemented or accepted socially or in a business. The Relevance dimension covers functionality and utility. Specificity covers how well the design description is complete and detailed, but Implicational Explicitness is defined as “The degree to which there is a clear relationship between the recommended action and the expected outcome”. This seems to apply much more towards creative ideas for business, than for design.

Unfortunately, scores for each factor are judged by humans, using tables of examples provided by the authors. Each factor score (e.g., 3, 2, 1) has a general description of what something of that level should possess, as well as some examples specific to the domain. The authors recommend setting thresholds if it is necessary to make a yes/no determination of Novelty for example. They suggest that summations of scores across dimensions allow “strength in one construct” to “compensate for weakness in another construct”. Deciding on thresholds, scores, and matching all make CDC use difficult.

2.4. Horn & Salvendy

Horn & Salvendy [2006] considers product creativity from the product consumer’s perspective, using Novelty, Resolution, Elaboration and Synthesis, from an older CPAM model, that have often been used to evaluate the “perception of product creativity”.

However, their scheme differs from many others by adding evaluations of **Affect** and **Preference**. Affect is the “emotional impact of product creativity”, defined in terms of Pleasure and Arousal. Although a number of studies have found that emotional impact has an important effect on the evaluation of creativity, it will be hard to determine the actual or potential emotional impact computationally with ease and accuracy.

Preference is the “preference for product creativity”, defined in terms of Centrality (“the consumer’s interest in creativity”) and Applicability (“the importance of creativity to the consumer”). Preference plays a role in reducing the effect of individual differences, much in the same way that Recommendation Systems use scoring schemes that account for individuals who consistently use evaluation scores in the range 1-5 as opposed to those who use scores in the range 4-7.

Horn & Salvendy experimentally reduced their scheme to six dimensions: Novelty, Resolution, Emotion, Centrality, Importance and Desire. Novelty and Resolution are from the CPAM, requiring originality and value. Emotion is emotional effect. The last three dimensions reflect the consumer’s individual preference for the product. Centrality is concerned with how the product matches the consumer’s interests; Importance is concerned with how relevant the product is to the consumer’s application; while Desire is about the “criticality” and “desirability” of the product.

Notice the large part that consumer behavior plays in this scheme, with the potential for use in marketing. Once again we note that this should cause problems for computational evaluation. However, that should not be grounds for being ignored.

2.5. Oman et al.

Oman et al. [2013] focus on creativity and innovation metrics, using Shah et al.’s [2003] ideas as a basis. Their Comparative Creativity Assessment (CCA) technique combines Novelty & Quality measures into one score: i.e., exactly what Shah says not to do! Their aim, however, is to reduce the amount of human judgment compared to the Shah’s approach. They want it to take less time, so the good news is that perhaps that might take less computation. To evaluate an individual design, it requires finding a comparison group of designs that address the same problem. Of course, that requires actually knowing which designs solve the same problem. It also suffers from most of the value setting and matching problems already mentioned.

Their main contribution, however, may well be the Multi-Point Creativity Assessment (MPCA) method, which is unusual because it takes into account which criteria each judge (of creativity) thinks is more important. Judges are given pair-wise comparisons between terms from the scales (the criteria) and the importance weights are taken to be the number of times each term in the pair is seen as more important. To model a particular judge, or a known population, it should be possible to factor out and precompute these weights for computational use of this method.

Each judge rates a product as scores from eight scales: Original – Unoriginal; Well-Made – Crude; Surprising – Expected; Ordered – Disordered; Astonishing – Common; Functional – Non-Functional; Unique – Ordinary; Logical – Illogical. There is no source given for these scales, but six of the eight dimensions can be easily mapped to dimensions used by others. However, the difference between Surprising and Astonishing isn't clear, nor is the difference between Unoriginal, Expected, Common and Ordinary. That is probably part of the reason for lower interrater reliability for the MPCA. As judgment of scores from scales is required, use in a CDC system would require an equivalent method to produce those scores.

2.6. Sarkar & Chakrabarti

Sarkar & Chakrabarti [2011] focus a lot on Novelty, comparing a product with "any other available product": allowing for a wide ranging comparison set [Brown 2014a]. They analyze a product for a new function ("very high novelty"). A new structure is certainly "novel", with subsequent analysis based on the "SAPPhIRE model of causality": novelty just based on parts is "low"; based on all aspects of the product including the state changes that present the external behavior ("high novelty"); and anything else is "medium" novelty. Note that this requires the comparison set to be stored as, or convertible to, something akin to a Structure, Behavior & Function representation. Also note that, as for other work, the novelty categorization, the associated scores for each category, and the actual scores used are all approximate with heuristic mappings involved.

Their other focus is on Usefulness, believing that to be creative a produce must be novel and useful. They ignore all the other aspects that we have discussed above. They argue that usefulness should correspond to actual use, but that importance is a factor. Importance can be scored as "life-saving" down to just "entertaining". Sarkar & Chakrabarti also suggest that if more people use a product it is more useful, and if the frequency and duration of use are high it is also more useful. While these could be assessed for existing types of products, if records exist, the more novel the product the harder it will be to correctly use historical data. Computers would have an even harder time.

In order to calculate product creativity, the novelty and usefulness scores are multiplied together. Their experimental evaluation suggests that Novelty plays a larger part in creativity evaluation than usefulness and that their proposed method matches "the intuition of experienced designers better than currently available methods” (to be fair, only two others were used, including Shah et al. [2003]).

3. Conclusions

An obvious contender for CDC system use, Shah et al. [2003] “Metrics for measuring ideation effectiveness”, has been discussed elsewhere as a possible guide for computational use [Brown 2014b]. Unfortunately, while very influential and demonstrably useful for design experiments, it has a number of problems that prevent it from being used computationally without modification.

Another contender is the series of papers by Maher, Fisher and co-authors (e.g., [Grace et al. 2014]) that address computational approaches to creativity evaluation. Theirs is probably the most automatable of the approaches in the research literature. However, while they rely on statistical methods (such as clustering and linear regression), by comparison with much of the other literature they lack experimental validation by testing with users, consumers or designers.

The most notable conclusion to take away from this short review is that in experiments, slightly different types of subjects confirm different subsets of creativity aspects as being good.

While almost everyone these days uses the model of creativity evaluation that has a small set of major dimensions with several aspects per dimension, the good news is that there are a lot of overlaps between the aspects chosen. The bad news is that nobody knows how to combine them, or even whether to combine them at all. The worst news is that most aspects do not come in computable form, despite being shown to be viable for creativity evaluation.

4. References

S.P. Besemer (2006) Creating products in the age of design. How to improve your new product ideas! *New Forums Press, Inc.*

D.C. Brown (2013) “Guiding Computational Design Creativity Research”*, Int. Jnl. of Design Creativity and Innovation 1(1).*

D.C. Brown (2014a) “Computational Design Creativity Evaluation”, *Proc. 6th Int. Conf. on Design Computing and Cognition (DCC’14)*, London, UK.

D.C. Brown (2014b) “Problems with the Calculation of Novelty Metrics”, *Proc. Design Creativity Workshop, 6th Int. Conf. on Design Computing and Cognition (DCC’14)*, London, UK.

D.H. Cropley & J.C. Kaufman (2012) “Measuring Functional Creativity: Non-Expert Raters and the Creative Solution Diagnosis Scale”, *Journal of Creative Behavior 46(2),* 119-137.

D.L. Dean, J.M. Hender, T.L. Rodgers & E.L. Santanen (2006) “Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation”, Jnl. Association for Information Systems 7(10).

K. Grace, M.L. Maher, D. Fisher & K. Brady (2014) “Modeling expectation for evaluating surprise in design creativity”, *Proc. 6th Int. Conf. on Design Computing and Cognition (DCC’14)*, London, UK.

D. Horn & G. Salvendy (2006) “Product creativity: conceptual model, measurement and characteristics”, *Theoretical Issues in Ergonomics Science 7(4)*, 395-412

S.K. Oman, I.Y. Tumer, K. Wood & C. Seepersad (2013) “A comparison of creativity and innovation metrics and sample validation through in-class design projects”, *Research in Engineering Design 24(1)*, 65-92.

P. Sarkar & A. Chakrabarti (2011) “Assessing design creativity”, *Design Studies*, 32, 348-383.

J.J. Shah, N. Vargas-Hernandez & S.M. Smith (2003) “Metrics for measuring ideation effectiveness”, *Design Studies 24,* 111-134.