Scarlett R. Miller[1]
Mem. ASME
School of Engineering Design,
The Pennsylvania State University,
213-P Hammond Building,
University Park, PA 16802-1401
e-mail: scarlettmiller@psu.edu

Samuel T. Hunter
Mem. ASME
Department of Psychology,
The Pennsylvania State University,
University Park, PA 16802-1401
e-mail: sth11@psu.edu

Elizabeth Starkey
Mem. ASME
School of Engineering Design,
The Pennsylvania State University,
University Park, PA 16802-1401
e-mail: ems413@psu.edu

Sharath Ramachandran
School of Engineering Design,
The Pennsylvania State University,
University Park, PA 16802-1401
e-mail: sharath@psu.edu

Faez Ahmed
Mem. ASME
Department of Mechanical Engineering,
Northwestern University,
Evanston, IL 02139
e-mail: faez@mit.edu

Mark Fuge
Mem. ASME
Department of Mechanical Engineering,
The University of Maryland,
College Park, MD 20742
e-mail: fuge@umd.edu

How Should We Measure Creativity in Engineering Design? A Comparison Between Social Science and Engineering Approaches

Design researchers have long sought to understand the mechanisms that support creative idea development. However, one of the key challenges faced by the design community is how to effectively measure the nebulous construct of creativity. The social science and engineering communities have adopted two vastly different approaches to solving this problem, both of which have been deployed throughout engineering design research. The goal of this paper was to compare and contrast these two approaches using design ratings of nearly 1000 engineering design ideas. The results of this study identify that while these two methods provide similar ratings of idea quality, there was a statistically significant negative relationship between these methods for ratings of idea novelty. In addition, the results show discrepancies in the reliability and consistency of global ratings of creativity. The results of this study guide the deployment of idea ratings in engineering design research and evidence.
[DOI: 10.1115/1.4049061]

Keywords: design process, design theory and methodology

1 Introduction

As research in the effectiveness of ideation techniques has increased in engineering design, it has the inherent challenge of measuring the nebulous construct of creativity [1]. Assessing creativity of ideas in terms of novelty and appropriateness (correct, useful, valuable, or meaningful) [2] is vital to the engineering design discipline for several key reasons. First, valid measurement helps researchers to determine which design methods help individuals or teams to generate creative ideas most effectively or prolifically [3]. Second, valid quantification of creative performance provides a means for the designers to properly assess the creativity of their own ideas in hopes of developing more innovative solutions [4,5].

Although there exists a plethora of metrics for measuring design creativity (see, e.g., Refs. [6–10]), these methods have been criticized for their lack of generalizability across domains [11], the subjectivity of the measurements [12], the vagueness of the measurement methods [13], and the timeliness of the method for evaluating numerous concepts [14]. There is also a lack of consistency across the literature and across disciplines for which creativity metric to use and when to use it. Because of this, design theory and methodology researchers have adopted a wide variety of metrics for assessing creativity including, but not limited to: the Consensual Assessment Technique (CAT) [15–19], expert panels [20–24], the Shah, Vargas-Hernandez, and Smith (SVS) method [3,25–30], SVS extensions [31,32], and other newly created metrics for creative design evaluation [30,33–38]. However, the two most widely adopted are the CAT and SVS methods (as well as its extensions).

The consensual assessment technique (CAT), put forth by Amabile [2,39,40], was developed by social scientists as a method for measuring creativity through subjective measures. It relies on the simple idea that an artifact is creative only to the extent to which "experts" in the area agree, independently, that it is creative. In contrast to this approach, the Shah, Vargas-Hernandez, and Smith (SVS) [3] method relies on breaking down design concepts into their components and then quantifying the creativity of the ideas based on relative frequencies.

One of the main issues with the adoption of these vastly different methods for measuring creativity is it can influence our ability to compare and contrast findings. This is particularly important

[1]Corresponding author.

Contributed by the Design Theory and Methodology Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received May 31, 2020; final manuscript received October 26, 2020; published online January 27, 2021. Assoc. Editor: Julie Linsey.

because recent research [41,42] has demonstrated that applying different creativity metrics to the same design problem can result in creativity rankings that are not only vastly different but often negatively correlated. This means that applying different metrics to the same design problem could result in research findings that contradict prior results on the sole basis of the creativity measure used in the study. However, these two widely adopted approaches (SVS and CAT) have yet to be compared making it unclear how, or if, research studies that have deployed these different approaches should be compared and contrasted.

Thus, the goal of the current study was to compare and contrast these two standard approaches by studying the creativity measurement of nearly 1000 design ideas generated by engineering design students. The results of this study can be used to inform how we apply and compare creativity results in engineering design research.

2 Related Work

Before we can begin to compare and contrast these two approaches to measuring creativity, it is first important to review the rationale for their creation and adoption in their respective fields. Thus, the current section serves to highlight research on creativity measurements in the social science and engineering disciplines that provide a groundwork for the current study.

2.1 A Social Science Approach to Creativity Measurement. The consensual assessment technique (CAT) [2,39,40] has been widely adopted by the social science community and is backed by over 30 years of research that has identified it as a reliable and valid way of measuring creativity. The method is grounded on the consensus of individuals with knowledge about a given domain, or "experts" (see discussions by Baer et al. [43]; Kaufman et al. [44]). This group of researchers contends that while creativity can be difficult to characterize in terms of specific features, it is something that people can recognize and agree upon when they see it. They also believe that creativity judgments can only be subjective, and researchers should not attempt to objectify the creative rating processes (see discussion in Ref. [45]).

In the CAT method, a panel of independent "expert" raters who are familiar with the domain and who have not conferred with one another are recruited and asked to independently make assessments of a product's creativity through the use of a Likert Scale. The specific dimensions of creativity can vary from a global assessment of creativity [46,47] to a series of sub-dimensions that comprise the construct in a given domain (e.g., Jeffries [48]). An often used taxonomy includes ratings product novelty (e.g., original or surprising), quality or utility of the product (e.g., valuable, logical, useful, and understandable), and product elegance (organic, well-crafted) [49].

As originally conceptualized, one of the central components of the CAT is the use of an appropriate group of judges to make the creativity assessments [50]. Specifically, Amabile [39,40] suggested that expertise within a given domain is necessary to make accurate assessments of creative products. As would be expected, numerous researchers have demonstrated that expert judges typically produce more similar ratings (higher interrater reliability) than non-expert raters (see, e.g., Refs. [39,40]). In addition, a more formal and larger-scale test of the role of expertise in assessing creativity was conducted by Kaufman and colleagues [51] who assessed the creativity of poems generated by college students. This study showed that experts, once again, produced stronger interrater reliability relative to novice judges who were less consistent in their agreement on creativity judgment. Moreover, the correlation between experts and non-experts was rather low ($r = 0.22$) suggesting that when rating more complex outcomes, experts and novices may be rating differing constructs. The extension of this reasoning is that as a product grows in complexity, the use of experts will become more important to producing accurate ratings. That is, the gap in creativity rating accuracy is likely to grow between experts and novices in complex domains like physics and engineering [50,52].

An extension of the above is the important caveat that in more simplistic domains or with less complex products, it may be possible for novices to approximate the ratings of experts. Indeed, in a study of the creativity of short stories, Kaufman and colleagues [53] concluded that the correlation of 0.89 between experts and novices was evidence that if enough novice raters are used, "they may be as reliable as experts." Moreover, some researchers have attempted to approximate expertise via the use of training techniques prior to ratings. Specifically, using the modified Q-sort technique [54], researchers ask knowledgeable individuals (i.e., experts) to select exemplars or benchmarks of what constitutes, for example, a highly creative product and a highly uncreative product. Using these exemplars, raters can be trained to produce ratings that approximate the mental model of expert ratings (e.g., Hunter et al. [55]; Lovelace and Hunter [56]).

Although it is possible, in some instances, for novices or trained novices to produce ratings commensurate with experts, within the domain of engineering and design, it remains open to question as to whether the complexity of the products being assessed allows novices to be reliably utilized. Importantly, as noted by Kaufman and Baer [50], "If non-experts and experts do not agree with each other, then the opinion of experts in a domain should trump those of anyone else." This means, if this finding holds true in the engineering domain, experts need to be solely used to judge the creativity of engineering products. However, those that have adopted this method in engineering research often rely on non-expert judges like third-year engineering students (see, e.g., Ref. [57]) due to the difficulty of finding experts to perform these evaluations. The use of novices or trained novices in these evaluations is often due to the time required to perform such evaluations. This brings to question if and when novices can be used to evaluate creativity metrics. While the need for identifying suitable judges was highlighted in recent critical evaluation of the CAT in the psychology literature [18], no study to date has explored the impact of expertise on the deployment of the CAT in an engineering context, thus leaving it unclear if these expertise gaps are apparent in the engineering domain. Thus, the current study seeks to fill this research void.

2.2 An Engineering Approach to Measuring "Ideation Effectiveness". In contrast to social science research, the majority of creativity research in engineering has focused on quantifiable measures of ideation method *effectiveness*. The term ideation effectiveness is often used in engineering research due to the "difficulty in defining this term (and agreeing on its meaning)" [3]. These metrics typically rely on breaking down design concepts into their components and then quantifying the creativity of each of these components by various means. Instead of measuring creativity, SVS proposed to study four metrics (quantity, quality, novelty, and variety) of *effectiveness*. Of these four metrics, quantity and variety measure ideation effectiveness holistically (at the idea set level) while novelty and quality can be measured at the individual idea level. Most central to the current discussion and comparison with the social science metrics for creativity are the calculations of the SVS novelty and quality metrics due to our adoption of the widely accepted definition of creativity as something that is both novel and appropriate [2] and our measurement of individual ideas rather than idea sets.

SVS defined quality as "a measure of the feasibility of an idea and how close it comes to meet the design specifications" [3]. SVS argued that an idea's quality can be measured as a physical property even at the conceptual stage where it can be adequately estimated even though there is not enough information to do quantitative analysis. They suggest that the technical feasibility of an idea can be evaluated using questions like "how fast can it go" or "can it get off the ground" through both experiential and analytical knowledge. While they propose to evaluate ideas using engineering analyses like quality function deployment (QFD) [58] or the Pugh Matrix [59], these

methods are difficult to employ for early-stage conceptual concepts. Instead, quality is often scored on these early-phase ideas by two raters who use a three- or four-point rating scale to evaluate the technical feasibility and difficulty of the design, see Ref. [60] for discussion. This multi-point scale was developed because prior work in engineering had shown that raters had difficulty applying an unanchored scale which led to low consistency between raters [61].

On the other hand, the SVS novelty metric is based on relative creativity, or "how unusual or unexpected an idea is compared to other ideas" [3]. The SVS approach relies on the development of a genealogy or feature tree to calculate the relative design novelty of an idea by identifying features like motion type and control mechanism and then the different ways in which each of those attributes is satisfied [3]. Concepts with features in categories with lower frequency counts are considered more novel, whereas designs with features with higher frequency counts are considered less novel because they occurred more frequently in the sample studied. This method has become widely adopted in engineering due to limited rater bias [3,62]. However, many limitations have been reported such as low interrater reliability, inaccurate representations, and difficulties interpreting multiple metrics simultaneously [39,40]. In addition, the use of the SVS method for large data sets is limited as differences in novelty values for large sets are diminished due to the relative nature of the metric [30].

Because of these pitfalls, a wealth of extensions to this metric have been proposed and implemented in engineering research [7–10]. For example, Hernandez et al. [8] took the genealogy tree approach developed for assessing the variety of ideas for an individual in the SVS metrics and decided to merge the individual trees to compose novelty scores over a data set. In addition, Peeters et al. [10] developed a method to look at three different levels of the novelty of an idea (physical principles, working principles, and embodiment) through a similar genealogy tree approach. While both of these metrics can broaden the range of novelty scores over the data set, they do not do well for incomplete ideas, or ideas that do not have an embodiment. Therefore, Johnson et al. [7] developed their new novelty metrics that will score ideas with or without embodiment level details, allowing the metric to support abstract responses. In addition, these new metrics allow for better control of edge cases [7].

3 Research Objectives

The evaluation presented in the current paper was developed through a discussion between the authors, a combination of engineers and a psychologist, when they debated which creativity metric to use to analyze their data for a design study. In light of these discussions, the following research questions were developed:

> RQ1: Do the gold standard metrics used in the social science and engineering disciplines measure the same construct of design novelty and quality?
>
> RQ2: Can trained novices be used as a proxy for experts when measuring subjective novelty and quality of an idea in the engineering design domain?
>
> RQ3: Can (or should) global assessments of creativity be used in the engineering design domain?

The remainder of this paper highlights the analysis and comparison of these two approaches, the evaluation of who does the evaluation, and their utility for design studies.

4 Previous Work

A prior research study was conducted with 141 engineering students (89 freshmen and 52 seniors; 95 males and 46 females) geared at identifying the influence of product dissection on engineering learning and creativity [63]. During this study, the participants were asked to complete a product dissection activity and then participate in a 20-min brainstorming activity where they sketched ideas for the following design prompt:

> Upper management has put your team in charge of developing a concept for a new innovative product that froths milk in a short amount of time. Frothed milk is a pourable, virtually liquid foam that tastes rich and sweet. It is an ingredient in many coffee beverages, especially espresso-based coffee drinks (Lattes, Cappuccinos, Mochas). Frothed milk is made by incorporating very small air bubbles throughout the entire body of the milk through some form of vigorous motion. The design you develop should be able to be used by the consumer with minimal instruction. It will be up to the board of directors to determine if your project will be carried on into production.

The participants in this prior study created a total of 932 concepts which included both visual images (sketches) and a short textual description of the idea, see Fig. 1 for example sketches.

4.1 Novelty and Quality Metrics. To investigate the influence of the creativity metrics used on measured creativity, the creativity of the 932 ideas were analyzed in four primary ways: (1) novelty and quality from experts using the social science approach of the CAT, (2) novelty and quality ratings from trained novices using the CAT method, (3) novelty and quality ratings from the assessors employing the engineering SVS method, and (4) novelty ratings from the assessors employing an extension of the SVS method [7]. These approaches are summarized in Table 1. The remainder of this section describes how novelty and quality were analyzed in the current study.

4.1.1 Consensual Assessment Technique Ratings. For both expert and non-experts, the guidelines put forth by Besemer [49] and Besemer and O'Quinn [64] were used, namely, raters (expert and novices) in approximately 20 h of training sessions which involved: a history of the Consensual Assessment Technique its history and use, the potential impact of bias in creativity ratings (e.g., central tendency bias and avoiding extremes), and a discussion of the meaning of the different ratings provided for each idea (novelty and quality). Finally, sample ideas were reviewed during the final training session, and the raters practiced providing novelty and quality ratings for these ideas and discussing the rationale for these ratings. Specifically, raters provided a rating from 1 (low novel or quality) to 7 (high novelty or quality) based on the quality definition of value, logic, utility, and how understandable the ideas were and the novelty definition of originality or surprise. Following the training sessions, raters performed the ratings for the 943 ideas for one month. Raters provided these assessments independently, and the scores were aggregated.

To justify the application of the expert label, one rater had graduate degrees and the other had completed graduate coursework, both in an engineering design-related field. In addition, both raters had at least four years of applied experience in both design and assessment and had published, minimally, six papers on the topics of design and creativity assessment. These experts were selected based on Amabile's suggestion that "expertise within a given domain is necessary to make accurate assessments of creative products" [39,40]. These expert ratings are the same used in the prior study [63]. On the other hand, three novice raters were recruited for the current paper to provide ratings of the same idea set. Specifically, three undergraduate psychology students with experience in coding and assessing creativity in at least three previous projects provided ratings for the idea set.

4.1.2 Shah, Vargas-Hernandez and Smith Ratings. SVS proposed two different approaches to measuring novelty [3], the first of which requires determining what concepts are not novel, while the second method, deployed here, requires researchers to measure the frequency with which a given idea is found in an idea set. Since SVS defines novelty as "how unusual or unexpected an idea is as compared to other ideas" [3], SVS-inspired methods

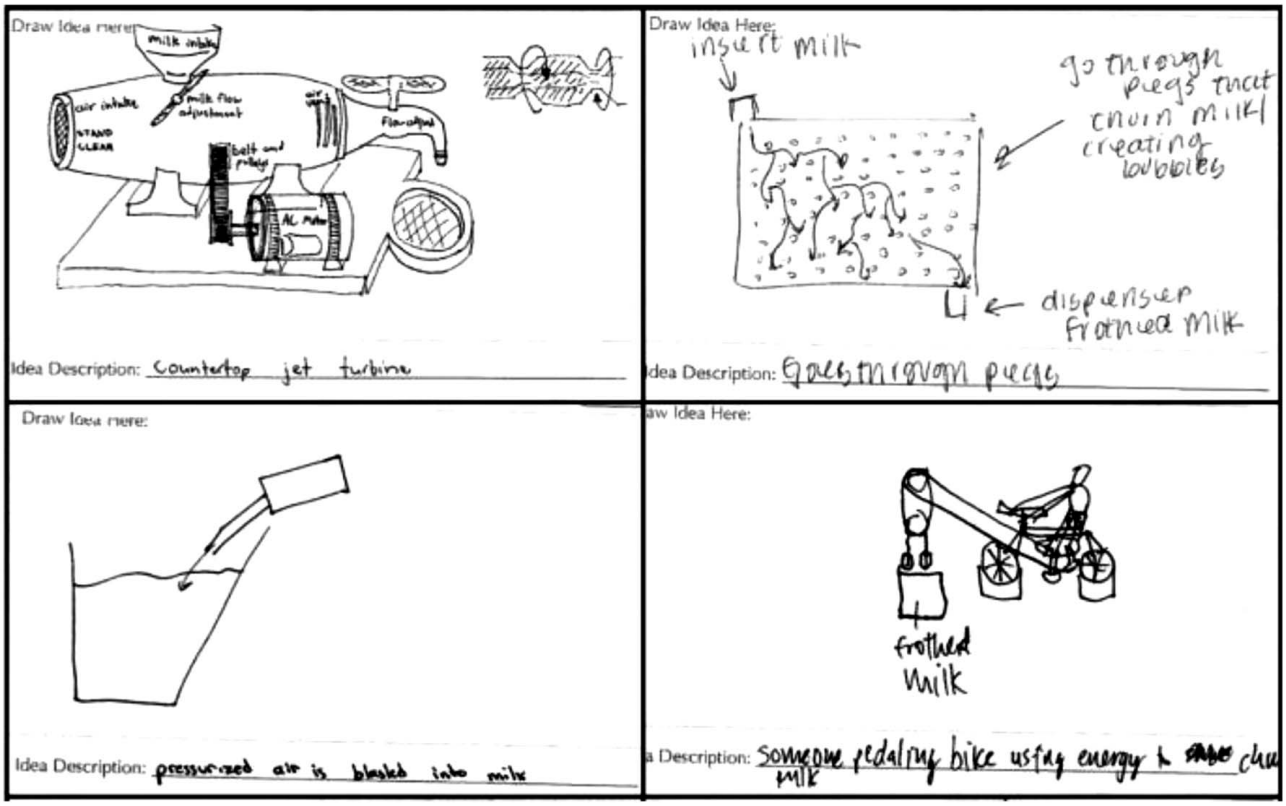

Fig. 1 Example of sketches provided to four expert raters during the qualitative study

Table 1 Summary of creativity ratings used in the current investigation

	Metrics	Rating method	Requires expert?	Non-expert training	Scale
CAT	Novelty and quality	Qualitative ratings	Yes[a]	~20 h	1–7
SVS	Novelty and quality	Feature tree	No	~10 h	0–1
Johnson et al.	Novelty	Feature tree	No	~10 h	0–10

[a]Non-experts can be used as a proxy of experts using a modified training technique.

generally look at novelty in a relative fashion, where concept novelty is compared to ideas from the same idea set. For the current analysis, the novelty of the ideas was calculated by identifying the novelty of each feature within the idea and then comparing these features to all of the designs being reviewed [3]. Ultimately, these calculations produce a value between 0 and 1. Designs with novelty values closer to 0 indicate less novel concepts while novelty values closer to 1 indicate concepts that are more novel.

In order to calculate design novelty, two raters, a graduate and a undergraduate student in engineering, were recruited. Prior to this assessment, the raters received extensive training on the design tasks and rating process. One of these raters was also an expert CAT rater in order to maintain consistency across ratings. However, it is important to note that the CAT ratings were done prior to the SVS ratings. In order to rate the designs, a Design Rating Survey (DRS) was used to help the raters classify the features' each design concept addressed as described in Ref. [3]. The DRS contained 24 questions for the Milk Frother design task; the first 20 questions on the DRS were used to help raters classify the features' each design concept addressed, similar to the feature tree approach used in previous studies to compute design novelty (see Refs. [65,66] and more details). The interrater agreement was 0.85 for this approach. The results from these concept evaluations were used to calculate the novelty of the generated ideas according to SVS [3] calculations through the process described in detail by Toh and Miller [67].

In addition to design novelty, SVS also defines design quality as "the feasibility of an idea, and how close it comes to meet the design specifications" [3]. In the current study, the quality values were calculated using the final four survey questions on the DRS designed according to the approach used by Linsey et al. [60]. These questions are as follows: (1) Will it froth milk? (2) Is it technically feasible to execute? (3) Is it technically easy to execute? and (4) Is it a significant improvement over the original design? Any disagreements were settled in a conference between the two raters. By answering these questions, the quality is evaluated on a 4-point scale that is normalized (by dividing the human responses by 10 to attain a score between 0 and 1 with 1 considered as the maximum absolute quality rating). The interrater agreement was 0.62 for this approach. The details of this calculation are described in the study by Toh and Miller [67].

4.1.3 Johnson et al.'s Novelty Metric (Extension of SVS Novelty). Johnson et al.'s [7] novelty metric was developed to extend the SVS approach to include ideas that are at higher levels of abstraction, to support changes in the SVS genealogy tree and to support changes in the data set in a meaningful way. In the current study, this metric was utilized to see if improvements to the SVS method resulted in improvements in the relationship between social science and engineering approaches to measure the creativity. In order to calculate this metric, the results from the previously developed DRS was used to classify the features

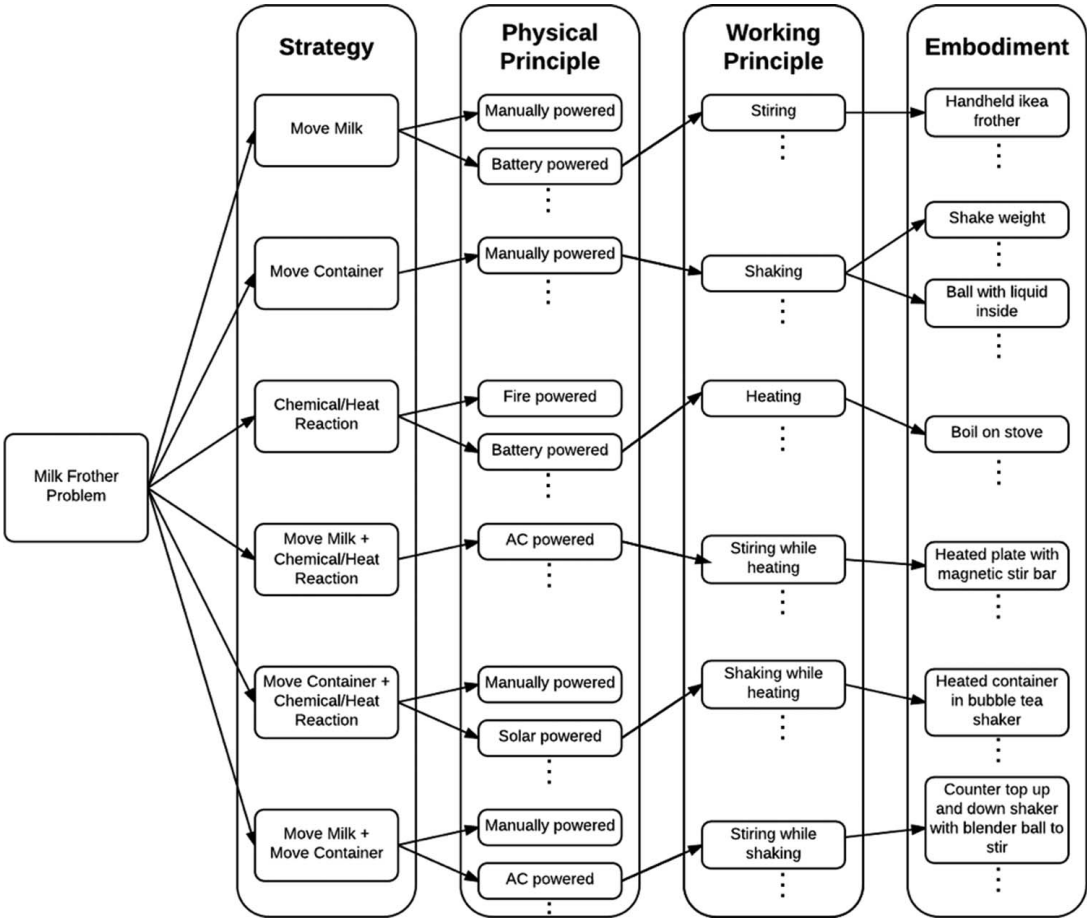

Fig. 2 Feature tree used to calculate the Johnson et al. (2016) novelty metric

addressed by each design concept. The results of the DRS were then split into which category they addressed in the extension metrics: strategy, physical principle, working principle, or embodiment (see Fig. 2 for details). The strategy was determined by how the product achieved the act of frothing (i.e., by moving the milk or moving the container with milk) while the physical principle was determined by what type of power source was used to power the product (i.e., manual, battery). On the other hand, the working principle was evaluated by determining what type of motion was used by the product (i.e., stirring, shaking) and the embodiment was determined by how the product looked like (i.e., shake weight, handheld frother). Total novelty scores were determined by the equations described in Ref. [7], using the weight of 10 for strategy, 6 for physical principle, 3 for working principle, and 1 for embodiment. These weights were selected as proposed in the initial paper [7] to mimic the weights Shah used for scoring variety of a genealogy tree.

5 Data Analysis and Results

In order to address our research goals, the novelty, quality, and general creativity of 932 concepts were assessed. Table 2 provides an overview of our results while the remainder of this section presents our results with reference to our research questions. SPSS v.24 was used to analyze the results, and a significant level of 0.05 was used in all analyses and effect sizes were classified according to Cohen [68].

5.1 RQ1: Do the Standard Metrics Used in the Social Science and Engineering Disciplines Measure the Same Construct of Design Novelty and Quality?. Our first research question was developed to understand if the standard creativity metrics used in the social sciences (CAT expert ratings) and the engineering domain (SVS and its extension) were measuring the same construct of creativity through novelty and quality assessments. The results revealed a lack of a strong relationship between the scores generated using the SVS method and its extension and those scores using the CAT method (see Table 2 for the full correlation results). In fact, expert novelty was *negatively* correlated with SVS ratings of novelty ($r = -0.11$, $p = 0.002$). While the extended SVS novelty metric by Johnson et al. [7] was found to be positively correlated with expert novelty ($r = .14$, $p < 0.001$), the effect was small. On the other hand, expert quality ratings were positively related to SVS quality ($r = .31$, $p < 0.001$), a medium effect size. The implication here is that there is a disconnect between the widely used and accepted methods of measuring design novelty in the social sciences (CAT) and engineering (SVS and its extensions) domains. On the other hand, the quality ratings, which were completed using a 4-point qualitative scale for the SVS method, seem to be guiding raters to measure similar constructs of quality, shown by the correlation between SVS and CAT expert quality ratings.

5.2 RQ2: Can Trained Novices Be Used as a Proxy for Experts When Measuring Subjective Novelty and Quality of an Idea in the Engineering Design Domain?. Given that the previous finding indicated differences between the engineering and social science approach to measure design novelty, our second research question sought to understand if novices could be used as a proxy for measuring subjective creativity (CAT) in the engineering domain. In order to examine this research question, we examined the degree to which both sets of raters (experts and

Table 2 Correlations among creativity outcomes

		Social science CAT ratings				Engineering metrics		
		Expert novelty	Expert quality	Trained novice novelty	Trained novice quality	SVS Novelty	SVS Quality	Johnson et al.'s. novelty
Social science CAT ratings	Expert novelty	(0.71)	–	–	–	–	–	–
	Expert quality	−0.29	(0.75)	–	–	–	–	–
	Trained novice novelty	0.74	−0.34	(0.78)	–	–	–	–
	Trained novice quality	−0.41	0.5	−0.5	(0.56)	–	–	–
Engineering metrics	SVS novelty	−0.1	0.3	−0.11	0.35	(0.85)	–	–
	SVS quality	−0.22	0.31	−0.3	0.32	0.17	(0.62)	–
	Johnson et al.'s novelty	0.14	0.09	0.09	0.06	0.39	0.17	(0.85)

Note: All bold correlations statistically significant at $p < .05$, $n = 932$ (ICC2 values, i.e., interrater reliability, values in brackets).

trained novices) provided similar values, also known as interrater reliability, for ratings made using the CAT. In order to determine this, intraclass correlation coefficient (ICC2) was used—as noted by LeBreton and Senter [69], ICC(2) represents an understanding of "the extent to which the mean rating assigned by a group of judges is reliable". The values range 0–1 with zero being low (or no) reliability and 1 being perfect reliability or consistency in ratings. Values above 0.7 are typically considered acceptable with regard to assessing consistency across judges (i.e., judges are rating things similarly) [70,71].

Using ICC2, we found that experts provided similar ratings to one another as depicted by meeting the threshold of 0.70 [70,71] for both novelty (ICC2 = 0.71) and quality (ICC2 = 0.75) assessments. While trained novices were able to provide ratings of sufficient similarity for ratings of novelty (ICC2 = 0.78), they were not for quality as agreement fell below the 0.70 ICC2 threshold (ICC2 = 0.56). Consistent with trends on the interrater reliability findings, correlations between the aggregated ratings of experts and aggregated ratings of trained novices were higher for assessments of novelty ($r = 0.74$, $p < 0.001$) than quality ($r = 0.50$, $p < 0.001$), see Table 2.

On the whole, these results suggest that although experts and trained novices are capable of providing consistent ratings of novelty, they are less consistent when assessing the quality. Although we cannot directly test accuracy the of quality ratings given the nature of the data gathered, guidelines put forth by researchers such as Amabile [2] and Kaufman et al. [44] would suggest that trained novice scores are less accurate than experts with regard to quality. Put another way when assessing more complex phenomena (i.e., those found in design and engineering), it seems that trained novices can provide accurate ratings on whether a product is novel but are less consistent and accurate at providing input that a given product is of high quality. This point underscores the importance of following recommendations by researchers such as Besemer and O'Quin [49] who suggest that creativity is a multidimensional construct, comprised minimally of novelty and quality.

5.3 RQ3. Can (or Should) Global Assessments of Creativity Be Used in the Engineering Design Domain?. The third research question was developed to understand the accuracy of *global* assessments of creativity in the engineering design domain. As a reminder, a global assessment is when someone is asked to provide an overall subjective rating of creativity as opposed to a rating of quality and a rating of novelty, as discussed in *RQ1*. Similar to *RQ2*, intraclass correlation coefficient (ICC2) was used to assess the degree to which experts and novices provided similar (i.e., ratings that are internally consistent) global creativity ratings. Using an intraclass correlation coefficient (ICC2), we found that experts fell below the 0.70 ICC2 threshold (ICC2 = 0.43). Trained novices, however, were able to provide ratings of sufficient similarity for ratings of creativity (ICC2 = 0.70). Consistent with trends on the interrater reliability findings, correlations between experts and trained novices, while significant, were of a small effect for creativity ratings ($r = 0.14$, $p < 0.001$). These results indicate that experts had difficulty in producing similar ratings of the global creativity construct and, when they did, these ratings were minimally related to ratings made by trained novices. Taken further, the inconsistency in these findings begs the question as to what, precisely, experts and trained novices were rating? In particular, we sought to explore if experts and trained novices differentially weighted novelty and quality assessments when making global ratings of creativity.

Thus, to understand if the experts and trained novice placed differential weights on idea novelty and quality in their overall assessments of creativity, two linear regression analyses were conducted (see Table 3). In the first, the expert ratings of global creativity were regressed onto expert quality and novelty ratings. Results reveal that 54% of the total variance in global creativity is accounted for by quality and novelty ratings. Of particular note is that experts produced standardized regression coefficient of $B = 0.59$ for ratings of quality and only $B = 0.34$ for novelty, suggesting that quality was utilized more heavily than novelty. In fact, an inspection of the changes in R^2 values indicates that the quality variable accounted for 37% of the unique variance in ratings of global creativity while novelty accounted for only 17%.

Contrasting the results of the expert raters was those of the trained novices. Using an analogous regression analysis, the global assessment of creativity was regressed onto ratings of novelty and quality for trained novices. The total amount of variance accounted for was 0.73 or 73% of the variance, notably more than the experts and likely due to the higher interrater reliability of the ratings. What is particularly interesting, however, was that trained novices placed greater weight on novelty ($B = 0.75$) than quality ($B = 0.22$). When inspecting changes in R^2, in fact, only 5.5% of the variance was attributed to quality while a sizable 67% of the variance was associated with novelty.

Table 3 Regression summary of two linear regression analyses for experts and trained novices

		B	Std. error	Beta	t	Sig.
Expert	(Constant)	−0.205	0.092	–	−2.229	0.026
	Novelty	0.586	0.018	0.735	31.797	0
	Quality	0.336	0.018	0.435	18.817	0
Trained novices	(Constant)	−0.218	0.101	–	−2.158	0.031
	Novelty	0.220	0.019	0.233	11.797	0
	Quality	0.752	0.016	0.944	47.834	0

In summary, then, experts placed much greater weight on quality in their global assessments of creativity than trained novices who relied almost solely on novelty to determine their global ratings of creativity. When paired with the lack of strong interrater reliability on expert assessments of quality, these results suggest that using global assessments of creativity may be problematic. Namely, if global assessments of creativity are utilized, it is unclear what is being measured. Instead, we suggest, consistent with work by Besemer and O'Quin [72], that more fine-grained ratings of quality and novelty be utilized in engineering design research.

6 Discussion

The goal of this study was to understand *what*, if any, differences exist between social science (CAT) and engineering (SVS and extensions) approaches to measuring design novelty and quality and *why* differences may occur. The results of the study were as follows:

- When comparing expert CAT and SVS ratings, there was a statistically significant negative relationship for design novelty, but a positive relationship for design quality.
- While there was a significant agreement between trained novice and expert CAT novelty ratings, there was no significant agreement between these raters for design quality.
- There was a lack of significant agreement on global ratings of creativity by experts. In addition, experts and trained novices varied in the weight they placed on novelty and quality in their global assessments of creativity.

So, what do these results mean? First, the results identify that there is a significant negative relationship between expert CAT and SVS novelty ratings. This result would caution authors when comparing results from one novelty assessment (e.g., CAT) with prior work that utilized a different novelty assessment (e.g., SVS). This is because differences in findings may be related to the novelty assessment being used rather than the variables of study in the investigation. This is particularly important in the area of design theory and methodology as there are a plethora of novelty assessments being deployed in design studies (see, e.g., Refs. [15–19] [3,20–37]).

The results in this paper also identified significant agreement between expert and trained novice novelty ratings. This is of use to the engineering design community because expert raters come at great costs—particularly with larger design studies that produce more than 1000 design ideas. Thus, the results support the use of trained novices for novelty assessments in engineering design research when a modified Q-sort technique [54] is used. This type of method allows raters (even those outside of the field as demonstrated here) to produce ratings that approximate the mental model of expert ratings [55,56]. On the contrary, the results point to the fact that trained novice CAT rating may not be reliable when assessing the quality of early-phase design ideas. Instead, a guided quality assessment, such as SVS quality, or expert CAT ratings should be utilized to assess conceptual idea quality. This is in line with prior work by Kaufman and Baer [50] who stated that "If non-experts and experts do not agree with each other, then the opinion of experts in a domain should trump those of anyone else." Another potential source of deviance in these two approaches may also be in the way SVS calculated novelty—Is it measuring the uniqueness of ideas as the CAT tries to capture, or rather is it purely measuring rarity? Moving forward, not only do we need to clarify methods and approaches used by varying disciplines, but we must also work to establish consistent language that has clear meaning across disciplines. This includes, for example, the term "originality" that is often used in the social sciences to mean rarity or uniqueness, while SVS uses the term "novelty."

Finally, the results caution the use of global assessments of creativity. The results highlight a lack of agreement on the creativity of ideas by design experts. In addition, the results identify discrepancies between experts and trained novices on the weight of an ideas novelty or quality in creativity assessments. As such, in line with recommendations by Besemer and O'Quin [72], global assessments of creativity should be avoided without substantive validation prior to deployment.

6.1 Which Method Should Be Used?. Given these results and the lack of convergence between these popular methods of creativity assessment, a natural question emerges: What method should be used? Perhaps a variant on this core would be: When should each method be used? Unfortunately, it is too early to provide an answer to such questions, and instead, several steps must be taken before doing so.

Consider the following metaphor often used in science, namely an unknown or unclear phenomenon depicted as an elephant [73]. One researcher may hold tightly onto the trunk, confidently describing it as such. Another researcher may grasp the leg, confidently describing it as such. The reality is that both scientists are holding an elephant, and they simply have not connected both components to see the larger picture. Both the social science (CAT) and engineering (SVS and extensions) represent components of creativity and, like the elephant's trunk and leg, are very clearly dissimilar to one another on the surface. To connect such methods, we need to understand each in greater detail. We need to understand where the leg is on the body and that can help us understand that it is used to bear weight. We need to understand how the trunk moves to understand it is used in feeding. We need to connect the components to the larger whole.

In non-metaphorical terms, building a deeper understanding of each method will require building an expanded nomological network. That is, linking the social science and engineering measures to known correlates of creativity. Building this pattern of results will provide the contextual background of construct validity or what is being measured by each method. Being an older method, it is not surprising that CAT has some of this nomological network established [18], but more work is needed connecting CAT to design and engineering correlates, directly. With this constellation of relationships in place, scholars will have a clearer picture of both SVS and CAT, paving the way for recommendations on when each method is of the greatest utility.

Building a nomological network, like establishing construct validity, is not a "completed/not-completed" dichotomy but rather a process with degrees of "doneness" [74]. We recommend the following steps as guidelines for promoting a useful nomological network: (1) measure known antecedents or contextual predictors of creativity such as autonomy, resource availability, and climate [75]; (2) measure individual differences also associated with creativity, including personality and risk-taking [76]; (3) quantify known outcomes also associated with creativity that are also used as direct or proximal indicators of creative performance such as patents, client satisfaction, sales, customer reviews, and funding received [77]; and (4) finally, include measures to that provide discriminant validity or evidence that the measure (i.e., CAT or SVS) is not tapping into constructs they should not be. This might be, for example, preferences for favorite flower. This list is not exhaustive by any means but provides the reader with a foundation to explore a nomological network surrounding both CAT and SVS. With such measures in place along with indicators of creative performance as quantified by CAT and SVS, it will be possible to examine the pattern of effects and relationships among measured variables. To the extent that a given measure is related to known indicators of creativity, and not to those it should not be, evidence for construct validity is (or is not) established.

7 Conclusion

While the results found here can help inform design studies, there are several limitations and areas for future work. First, while the problem explored here was relatively simple, the results are likely to be exasperated in more complex problems like those found in

engineering design and systems engineering [50,52]. However, we do not yet know or fully understand what level of detail or complexity in a design task or ideation set is appropriate for use by the CAT and SVS methods. Further work is needed to identify if the results of this work will hold true for more complex or detailed level concepts by exploring a larger problem set.

In addition, given the importance of expertise in the rating process [39,40,51] and the findings of the study that clearly identify the difference between expert and trained novice raters in engineering design quality ratings, it is important to explore training methods for improving the viability and utility of rating assessments. This is particularly important in engineering due to the use of novices or trained novices in published articles (see, e.g., Ref. [57]), the difficulty in quantifying expertise in engineering domains which are multi-disciplinary in nature, and the time required by experts to perform these assessments (which often makes expert ratings unattainable).

Finally, creativity has several disputed definitions based on the domain, the environment, and the processes involved. In fact, engineering researchers have often strayed from using the word "creativity" and instead turn to the word "ideation effectiveness" due to "difficulty in defining this term (and agreeing on its meaning)" [3]. While our study is an example of how we can bring together multiple schools of thought in order to leverage the sought-out qualities from different metrics to create a more rigorous tool to quantify the abstract construct of creativity, future work is still needed. Specifically, while this paper highlights significant differences in how CAT and SVS rate design novelty, it does not provide finite evidence of "which metric is better" or if either method is truly appropriate in engineering design studies. Instead, it provides scientific evidence that key differences exist in these metrics that warrant further investigation. As such, future work should be on solving the wicked problem of arriving at meaningful methods to quantify nebulous constructs, such as creativity, that have multiple roots of origin and murky ground truths, which are currently debated across different domains.

Acknowledgment

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

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The data sets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request. The authors attest that all data for this study are included in the paper. Data provided by a third party are listed in Acknowledgment. No data, models, or codes were generated or used for this paper.

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," Proceedings of International Conference on Human Centered Design, Orlando, FL, July 9–14, Springer, pp. 120–129.
[2] Amabile, T., 1996, *Creativity in Context*, Westview Press, Boulder, CO.
[3] Shah, J. J., Smith, S. M., and Vargas-Hernandez, N., 2003, "Metrics for Measuring Ideation Effectiveness," Des. Studies, **24**(2), pp. 111–124.
[4] Christiaans, H. H., 2002, "Creativity as a Design Criterion," Commun. Res. J., **14**(1), pp. 41–54.
[5] Eshun, E. F., and de Graft-Johnson, K., 2012, "Learner Perceptions of Assessment of Creative Products in Communication Design," Art. Des. Commun. Higher Education, **10**(1), pp. 89–102.
[6] Borgianni, Y., Cascini, G., and Rotini, F., 2013, "Assessing Creativity of Design Projects: Criteria for the Service Engineering Field," Int. J. Des. Creativity Innovation, **1**(3), pp. 131–159.
[7] Johnson, T. A., Caldwell, B. W., Cheeley, A., and Green, M. G., 2016, "Comparison and Extension of Novelty Metrics for Problem-Solving Tasks," Proceedings of ASME 2016 International Design Engineering Technical Conferences & Computers and Information Engineering Conference, Charlotte, NC, Aug. 21–24, ASME, pp. 1–12.
[8] Hernandez, N., Okudan Kremer, G., and Schmidt, L. C., 2012, "Effectiveness Metrics for Ideation: Merging Genealogy Trees and Improving Novelty Metric," International Design Engineering Technical Conferences, Chicago, IL, Aug. 12–15, pp. 85–93.
[9] Nelson, B. A., Wilson, J. O., Rosen, D., and Yen, J., 2009, "Refined Metrics for Measuring Ideation Effectiveness," Des. Studies, **30**(6), pp. 737–743.
[10] Peeters, J., Verhaegen, P.-A., Vandevenne, D., and Duflou, J., 2010, "Refined Metrics for Measuring Novelty in Ideation," Proceedings of IDMME Virtual Concept 2010, Bordeaux, France, Oct. 20–22, pp 1–4.
[11] Baer, J., 2012, "Domain Specificity and the Limits of Creativity Theory," J. Creative Behav., **46**(1), pp. 16–29.
[12] Casakin, H., and Kreitler, S., 2005, "The Nature of Creativity in Design," Studying Des., **5**, pp. 87–100.
[13] Williams, A. P., Ostwald, M. J., and Askland, H. H., 2011, "The Relationship Between Creativity and Design and Its Implication for Design Education," Des. Principles Practice: An Int. J., **5**(1), pp. 57–72.
[14] Gosnell, C. A., and Miller, S. R., 2014, "A Novel Method for Assessing Design Concept Creativity Using Single-Word Adjectives and Semantic Similarity," ASME Design Engineering Technical Conferences, Buffalo, NY.
[15] 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," Des. Studies, **46**, pp. 6–37.
[16] Nikander, J. B., Liikkanen, L. A., and Laakso, M., 2014, "The Preference Effect in Design Concept Evaluation," Des. Studies, **35**(5), pp. 473–499.
[17] Baer, J., and Kaufman, J. C., 2019, "Assessing Creativity with the Consensual Assessment Technique," *The Palgrave Handbook of Social Creativity Research*, L. Izabela, and G. Vlad Petre, ed., Springer, New York, pp. 27–37.
[18] Cseh, G. M., and Jeffries, K. K., 2019, "A Scattered CAT: A Critical Evaluation of the Consensual Assessment Technique for Creativity Research," Psychol. Aesthetics, Creativity, Arts, **13**(2), p. 159.
[19] Stefanic, N., and Randles, C., 2015, "Examining the Reliability of Scores From the Consensual Assessment Technique in the Measurement of Individual and Small Group Creativity," Music Education Res., **17**(3), pp. 278–295.
[20] Alipour, L., Faizi, M., Moradi, A. M., and Akrami, G., 2017, "The Impact of Designers' Goals on Design-by-Analogy," Des. Studies, **51**, pp. 1–24.
[21] Cheng, P., Mugge, R., and Schoormans, J. P., 2014, "A New Strategy to Reduce Design Fixation: Presenting Partial Photographs to Designers," Des. Studies, **35**(4), pp. 374–391.
[22] Chan, J., Dow, S. P., and Schunn, C. D., 2015, "Do the Best Design Ideas (Really) Come From Conceptually Distant Sources of Inspiration?," Des. Studies, **36**, pp. 31–58.
[23] Baer, J., 2015, "The Importance of Domain-Specific Expertise in Creativity," Roeper Rev., **37**(3), pp. 165–178.
[24] Galati, F., 2015, "Complexity of Judgment: What Makes Possible the Convergence of Expert and Nonexpert Ratings in Assessing Creativity," Creativity Res. J., **27**(1), pp. 24–30.
[25] Vandevenne, D., Pieters, T., and Duflou, J., 2016, "Enhancing Novelty With Knowledge-Based Support for Biologically-Inspired Design," Des. Studies, **46**, pp. 152–173.
[26] Atilola, O., Tomko, M., and Linsey, J. S., 2016, "The Effects of Representation on Idea Generation and Design Fixation: A Study Comparing Sketches and Function Trees," Des. Studies, **42**, pp. 110–136.
[27] Toh, C. A., and Miller, S. R., 2015, "How Engineering Teams Select Design Concepts: A View Through the Lens of Creativity," Des. Studies, **38**, pp. 111–138.
[28] Tsenn, J., Atilola, O., McAdams, D. A., and Linsey, J. S., 2014, "The Effects of Time and Incubation on Design Concept Generation," Des. Studies, **35**(5), pp. 500–526.
[29] Doboli, A., and Umbarkar, A., 2014, "The Role of Precedents in Increasing Creativity During Iterative Design of Electronic Embedded Systems," Des. Studies, **35**(3), pp. 298–326.
[30] Sluis-Thiescheffer, W., Bekker, T., Eggen, B., Vermeeren, A., and De Ridder, H., 2016, "Measuring and Comparing Novelty for Design Solutions Generated by Young Children Through Different Design Methods," Des. Studies, **43**, pp. 48–73.
[31] Doboli, A., Umbarkar, A., Subramanian, V., and Doboli, S., 2014, "Two Experimental Studies on Creative Concept Combinations in Modular Design of Electronic Embedded Systems," Des. Studies, **35**(1), pp. 80–109.
[32] Starkey, E., Toh, C. A., and Miller, S. R., 2016, "Abandoning Creativity: The Evolution of Creative Ideas in Engineering Design Course Projects," Design Studies, **47**, pp. 47–72.
[33] Moreno, D. P., Hernandez, A. A., Yang, M. C., Otto, K. N., Hölttä-Otto, K., Linsey, J. S., Wood, K. L., and Linden, A., 2014, "Fundamental Studies in Design-by-Analogy: A Focus on Domain-Knowledge Experts and Applications to Transactional Design Problems," Des. Studies, **35**(3), pp. 232–272.

[34] Liu, W., Tan, R., Cao, G., Zhang, Z., Huang, S., and Liu, L., 2019, "A Proposed Radicality Evaluation Method for Design Ideas at Conceptual Design Stage," Comput. Ind. Eng., **132**, pp. 141–152.
[35] Christensen, B. T., and Ball, L. J., 2016, "Dimensions of Creative Evaluation: Distinct Design and Reasoning Strategies for Aesthetic, Functional and Originality Judgments," Des. Studies, **45**, pp. 116–136.
[36] Fischer, S., Oget, D., and Cavallucci, D., 2016, "The Evaluation of Creativity From the Perspective of Subject Matter and Training in Higher Education: Issues, Constraints and Limitations," Thinking Skills Creativity, **19**, pp. 123–135.
[37] Kershaw, T. C., Bhowmick, S., Seepersad, C. C., and Hölttä-Otto, K., 2019, "A Decision Tree Based Methodology for Evaluating Creativity in Engineering Design," Front. Psychol., **10**, p. 32.
[38] Goucher-Lambert, K., Gyory, J. T., Kotovsky, K., and Cagan, J., 2020, "Adaptive Inspirational Design Stimuli: Using Design Output to Computationally Search for Stimuli That Impact Concept Generation," ASME J. Mech. Des., **142**(9), p. 091401.
[39] Amabile, T. M., 1982, "Social Psychology of Creativity: A Consensual Assessment Technique," J. Personality Social Psychol., **43**(5), pp. 997–1013.
[40] Amabile, T., 1983, "Brilliant but Cruel: Perceptions of Negative Evaluators," J. Experimental Psychol., **19**(2), pp. 146–156.
[41] Gosnell, C. A., and Miller, S. R., 2016, "But Is It Creative? Delineating the Impact of Expertise and Concept Ratings on Creative Concept Selection," ASME J. Mech. Des., **138**(2), p. 021101.
[42] Sarkar, P., and Chakrabarti, A., 2011, "Assessing Design Creativity," Des. Studies, **32**(4), pp. 348–383.
[43] Baer, J., Kaufman, J. C., and Gentile, C. A., 2004, "Extension of the Consensual Assessment Technique to Nonparallel Creative Products," Creativity Res. J., **16**(1), pp. 113–117.
[44] Kaufman, J. C., Baer, J., Agars, M. D., and Loomis, D., 2010, "Creativity Stereotypes and the Consensual Assessment Technique," Creativity Res. J., **22**(2), pp. 200–205.
[45] Hennessey, B. A., Amabile, T. M., and Mueller, J. S., 1999, "Consensual Assessment," Encyclopedia Creativity, **1**, pp. 346–359.
[46] Cropley, D. H., Kaufman, J. C., and Cropley, A. J., 2011, "Measuring Creativity for Innovation Management," J. Technol. Manage. Innovation, **6**(3), pp. 13–30.
[47] Horn, D., and Salvendy, G., 2009, "Measuring Consumer Perception of Product Creativity: Impact on Satisfaction and Purchasability," Human Factors Ergonomics Manuf. Service Ind., **19**(3), pp. 223–240.
[48] Jeffries, K. K., 2017, "A CAT With Caveats: Is the Consensual Assessment Technique a Reliable Measure of Graphic Design Creativity?," Int. J. Des. Creativity Innovation, **5**(1–2), pp. 16–28.
[49] Bessemer, S. P., 1998, "Creative Product Analysis Matrix: Testing the Model Structure and a Comparison Among Products-Three Novel Chairs," Creativity Res. J., **11**(4), pp. 333–346.
[50] Kaufman, J. C., and Baer, J., 2012, "Beyond New and Appropriate: Who Decides What Is Creative?," Creativity Res. J., **24**(1), pp. 83–91.
[51] Kaufman, J. C., Baer, J., Cole, J. C., and Sexton∗, J. D., 2008, "A Comparison of Expert and Nonexpert Raters Using the Consensual Assessment Technique," Creativity Res. J., **20**(2), pp. 171–178.
[52] Hennessey, B. A., 1994, "The Consensual Assessment Technique: An Examination of the Relationship Between Ratings of Product and Process Creativity," Creativity Res. J., **7**(2), pp. 193–208.
[53] Kaufman, J. C., Baer, J., Cropley, D. H., Reiter-Palmon, R., and Sinnett, S., 2013, "Furious Activity vs. Understanding: How Much Expertise is Needed to Evaluate Creative Work?," Psychol. Aesthetics, Creativity, Arts, **7**(4), p. 332.
[54] Redmond, M. R., Mumford, M. D., and Teach, R., 1993, "Putting Creativity to Work: Effects of Leader Behavior on Subordinate Creativity," Org. Behav. Human Decision Processes, **55**(1), pp. 120–151.
[55] Hunter, S. T., Bedell-Avers, K. E., Hunsicker, C. M., Mumford, M. D., and Ligon, G. S., 2008, "Applying Multiple Knowledge Structures in Creative Thought: Effects on Idea Generation and Problem-Solving," Creativity Res. J., **20**(2), pp. 137–154.
[56] Lovelace, J. B., and Hunter, S. T., 2013, "Charismatic, Ideological, and Pragmatic Leaders' Influence on Subordinate Creative Performance Across the Creative Process," Creativity Res. J., **25**(1), pp. 59–74.
[57] Daly, S. R., Seifert, C. M., Yilmaz, S., and Gonzalez, R., 2016, "Comparing Ideation Techniques for Beginning Designers," ASME J. Mech. Des., **138**(10), p. 101108.
[58] Ter Harr, S., Clausling, D., and Eppinger, S., 1993, *Integration of Quality Function Deployment in the Design Structure Matrix*, Laboratory for Manufacturing and Productivity, Massachusetts Institute of Technology, Cambridge, MA, Working Paper No. LMP-93-004.
[59] Pugh, S., 1991, *Total Design*, Addison-Wesley, Reading, MA.
[60] Linsey, J. S., Clauss, E. F., Kurtoglu, T., Murphy, J. T., Wood, K. L., and Markman, A. B., 2011, "An Experimental Study of Group Idea Generation Techniques: Understanding the Roles of Idea Representation and Viewing Methods," ASME J. Mech. Des., **133**(3), p. 031008.
[61] Kurtoglu, T., Campbell, M. I., and Linsey, J. S., 2009, "An Experimental Study on the Effects of a Computational Design Tool on Concept Generation," Des. Studies, **30**(6), pp. 676–703.
[62] Oman, S. K., Tumer, I. Y., Wood, K., and Seepersad, C., 2013, "A Comparison of Creativity and Innovation Metrics and Sample Validation Through In-Class Design Projects," Res. Eng. Des., **24**(1), pp. 65–92.
[63] Starkey, E. M., Hunter, S. T., and Miller, S. R., 2019, "Are Creativity and Self-Efficacy at Odds? An Exploration in Variations of Product Dissection in Engineering Education," ASME J. Mech. Des., **141**(1), p. 012001.
[64] Besemer, S. P., and O'Quin, K., 1999, "Confirming the Three-Factor Creative Product Analysis Matrix Model in an American Sample," Creativity Res. J., **12**(4), pp. 287–296.
[65] Toh, C., and Miller, S., 2014, "The Role of Individual Risk Attitudes on the Selection of Creative Concepts in Engineering Design," ASME Design Engineering Technical Conferences, Buffalo, NY, Aug. 17–20, p. V007T07A027.
[66] Toh, C. A., and Miller, S. R., 2014, "The Impact of Example Modality and Physical Interactions on Design Creativity," ASME J. Mech. Des., **136**(9), p. 091004.
[67] Toh, C. A., and Miller, S. R., 2016, "Choosing Creativity: The Role of Individual Risk and Ambiguity Aversion on Creative Concept Selection in Engineering Design," Res. Eng. Des., **27**(3), pp. 195–219.
[68] Cohen, J., 1988, *The t Test for Means. Statistical Power Analysis for the Behavioural Sciences*, Lawrence Erlbaum Associates, Hillsdale, NJ.
[69] LeBreton, J. M., and Senter, J. L., 2008, "Answers to 20 Questions About Interrater Reliability and Interrater Agreement," Org. Res. Meth., **11**(4), pp. 815–852.
[70] Lance, C. E., Butts, M. M., and Michels, L. C., 2006, "The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say?," Org. Res. Meth., **9**(2), pp. 202–220.
[71] James, L. R., Demaree, R. G., and Wolf, G., 1984, "Estimating Within-Group Interrater Reliability With and Without Response Bias," J. Appl. Psychol., **69**(1), p. 85.
[72] Besemer, S., and O'Quin, K., 1986, "Analyzing Creative Products: Refinement and Test of a Judging Instrument," J. Creative Behav., **20**(2), pp. 115–126.
[73] Case, B., 1994, "Walking Around the Elephant: A Critical-Thinking Strategy for Decision Making," J. Continuing Education Nursing, **25**(3), pp. 101–109.
[74] Cronbach, L. J., and Meehl, P. E., 1955, "Construct Validity in Psychological Tests," Psychol. Bull., **52**(4), p. 281.
[75] Ma, H.-H., 2009, "The Effect Size of Variables Associated With Creativity: A Meta-analysis," Creativity Res. J., **21**(1), pp. 30–42.
[76] Toh, C., and Miller, S. R., 2019, "Does the Preferences for Creativity Scale Predict Engineering Students' Ability to Generate and Select Creative Design Alternatives?," ASME J. Mech. Des., **141**(6), p. 062001.
[77] Oldham, G. R., and Cummings, A., 1996, "Employee Creativity: Personal and Contextual Factors at Work," Acad. Manage. J., **39**(3), pp. 607–634.