Proceedings of the ASME 2020
International Design Engineering Technical Conferences
and Computers and Information in Engineering Conference
IDETC/CIE2020
August 17-19, 2020, Virtual, Online

DETC2020-22446

HOW SHOULD WE MEASURE CREATIVITY IN DESIGN STUDIES? A COMPARISON OF SOCIAL SCIENCE AND ENGINEERING APPROACHES

Scarlett R. Miller
School of Engineering Design,
The Pennsylvania State University,
University Park, PA, USA
Email: scarlettmiller@psu.edu

Samuel T. Hunter
Dept. of Psychology
The Pennsylvania State University,
University Park, PA, USA
Email: hunter@psu.edu

Elizabeth Starkey
School of Engineering Design,
The Pennsylvania State University,
University Park, PA, USA
Email: ems413@psu.edu

Sharath Ramachandran
School of Engineering Design,
The Pennsylvania State University,
University Park, PA, USA
Email: sharath@psu.edu

Faez Ahmed
Mechanical Engineering
Northwestern University
Evanston, IL
Email: faez00@umd.edu

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

ABSTRACT

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 paired with a qualitative study with expert raters. 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. Qualitative analysis of recordings from expert raters' think aloud concept mapping points to potential sources of disagreement. In addition, the results show that while quasi-expert and expert raters provided similar ratings of design novelty, there was not significant agreement between these groups for ratings of design quality. The results of this study provide guidance for the deployment of idea ratings in engineering design research and evidence for the development and potential modification of engineering design creativity metrics.

Keywords: design process, design theory, design theory and methodology

1. INTRODUCTION

As research in the effectiveness of ideation techniques has increased in engineering design research, so 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 determine which design methods help individuals or teams generate creative ideas most effectively or prolifically [3]. Second, valid quantification of creative performance provides a means for 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 for example [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 sciences as a 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 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 over 900 design ideas generated by engineering design students and identify potential causal factors of any discrepancies. The results from 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 discussion in Baer, et al., 2004 [43]; Kaufman, et al., 2010 [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 judgements can only be subjective, and researchers should not attempt to objectify the creative ratings process (see discussion in [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 (see Cropley, Kaufman, & Cropley, 2011 [46]; Horn & Salvendy, 2009 [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 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 inter-rater reliability) than non-expert raters (see for example [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 (2008) [51] who assessed the creativity of poems generated by college students. This study showed that experts, once again, produced stronger inter-rater 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 ($r = .22$) 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 .89 between experts and novices was evidence that if enough novice raters are used, "they may be as reliable as experts" (pg. 335). Moreover, some researchers have attempted to approximate expertise via the use of training techniques prior to ratings. Specifically, using the modified Q-sort technique (Redmond, Mumford & Teach, 1993 [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 & Hunter [56]).

Although it is possible, in some instances, for novices or quasi-experts 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" (pg. 85). This means, if these 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 3rd year engineering students (see for example [57]) due to the difficulty finding experts to perform these evaluations. The use of novices, or quasi-experts 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 this expertise gaps is apparent in the engineering domain. Thus, the current study seeks to fill this research void.

2.2 An Engineering Approach to Creativity Measurement

In contrast to social science research, the majority of creativity research in engineering has focused on quantifiable measures of an ideation methods *effectiveness*. The ideation effectiveness is often used in engineering research due to the *"difficulty in defining this term (and agreeing on its meaning)"* (pg. 116 [58]). 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 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" (pg. 117 [3]). They 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 QFD [59] or the Pugh Matrix [60], 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 that asks them to evaluate the technical feasibility and difficulty of the design, see [61] 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 [62].

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" (pg. 117 [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, 63]. However, many limitations have been reported such as low inter-rater reliability, inaccurate representations, and difficulties interpreting multiple metrics simultaneously [39, 40]. In addition, the use of the SVS method for large data sets has been found to be limited as differences in novelty values for large sets is 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, Okudan Kremer, and Schmidt [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 to help future design researchers appropriately understand *what*, if any, differences exist between social science and engineering approaches to measuring design novelty and quality and *why* these differences occur:

RQ1: Do the gold standard metrics used in the social science and engineering disciplines measure the same construct of design novelty and quality?

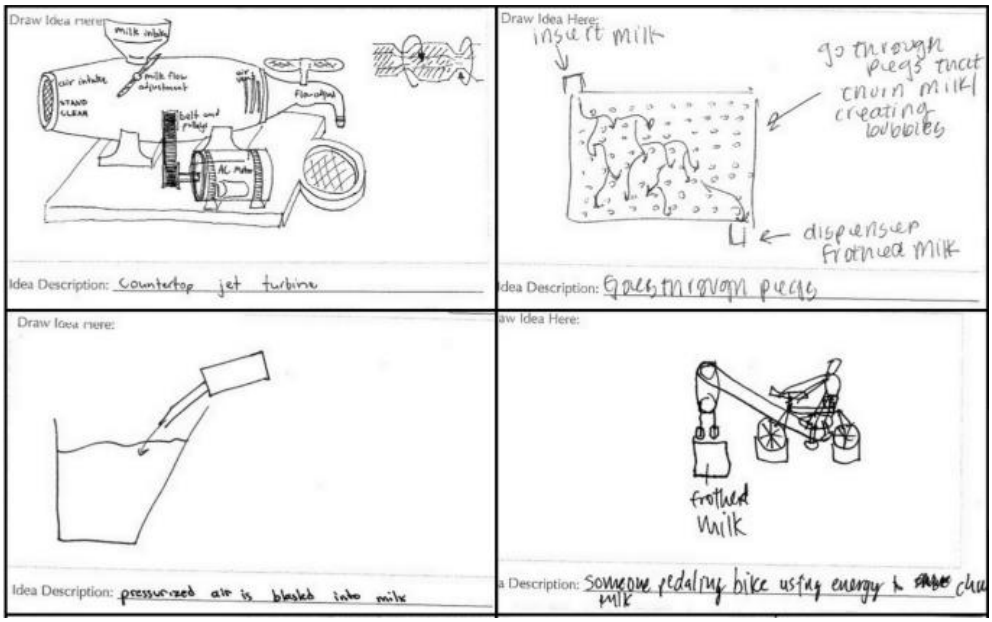

FIGURE 1: EXAMPLE OF SKETCHES PROVIDED TO FOUR EXPERT RATERS DURING QUALITATIVE STUDY.

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: What factors do human raters use to evaluate design novelty? How does this compare to the factors traditional used in engineering design research?

The remainder of this paper highlights the analysis and comparison of these two approaches and the factors that may contribute to similarities or differences in their measurement.

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. During this study, the participants were asked to complete a product dissection activity and then participate in a 20-minute 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) as well as a short textual description of the idea, see Figure 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 Consensual Assessment Technique (CAT), 2) novelty and quality ratings from quasi-experts 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.

TABLE 1: SUMMARY OF CREATIVITY RATINGS USED IN CURRENT INVESTIGATION. *NON-EXPERTS CAN BE USED AS A PROXY OF EXPERTS USING A MODIFIED TRAIING TECHNIQUE.

	Metrics	Rating Method	Requires expert?	Non-expert training	Scale
CAT	novelty & quality	Qualitative Ratings	Yes*	~ 20 hrs	1-7
SVS	novelty & quality	Feature Tree	No	~10 hrs	0-1
Johnson et. al	novelty	Feature Tree	No	~ 10 hrs	0-10

4.1.1 Consensual Assessment Technique (CAT) Ratings: For both expert and non-experts the guidelines put forth by Besemer [64] and Besemer and O'Quinn [65] were used. Namely, raters were asked to provide novelty metrics based on the definition of novelty as original and surprising. Raters provided quality scores using the definition of value, logic, utility, and how understandable the ideas were. Specifically, raters provided a rating from 1 (low novel or quality) to 7 (high novelty or quality). Raters provided these assessments independently and scores were aggregated. All ratings (both expert and non-expert) were completed over the course of one month.

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 in 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]." On the other hand, quasi-expert raters were undergraduate psychology students with experience coding and assessing creativity in at least three previous projects. In addition, the quasi-expert raters engaged in a minimum of 20 hours of rater training prior to providing ratings on the current project.

4.1.2 Shah, Vargas-Hernandez and Smith (SVS) 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" (pg. 117 [3]), SVS-inspired methods generally look at novelty in a relative fashion, where concept novelty is compared to ideas from the same idea set. For the current analysis, novelty was calculated based on the novelty of each feature within a design in comparison to the features within 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 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 the same as the CAT ratings in order to maintain consistency. 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 [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 [66, 67] and more details). The inter-rater 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 in Toh and Miller [68].

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" (pg. 117 [3]). In the current study, the quality values were calculated using the final 4 survey questions on the DRS designed according to the approach used by Linsey et al. [69]. These questions included: (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, 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 the maximum absolute quality rating. The inter-rater agreement was 0.62 for this approach. The details of this calculation are described in detailed in Toh and Miller [68].

4.1.3 Johnson et al. Novelty Metric (Extension of SVS novelty)

The Johnson et al. [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 dataset in a meaningful way. In the current study, this metric was utilized to see if improvements to the SVS method resulted in improvements between the relationship between the social science and engineering approaches to measuring creativity. In order to calculate this metric, the results from the previously developed Design Rating Survey (DRS) was used to classify the features addressed by each design concept. The results of the DRS were then split into which category they addressed in the extension metrics: strategy,

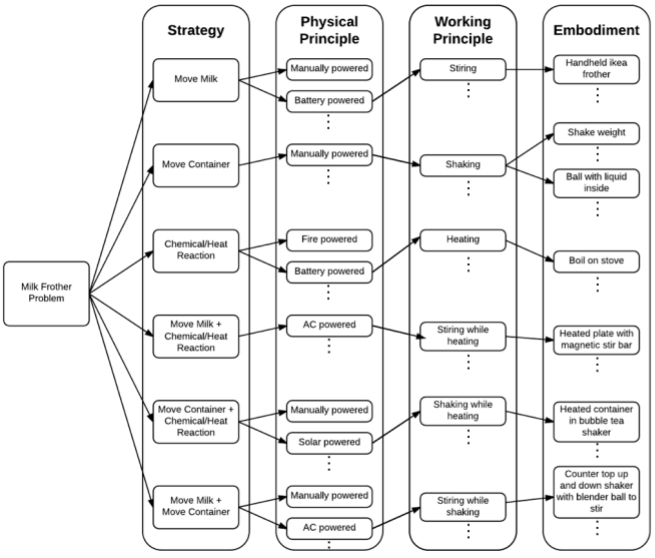

FIGURE 2: FEATURE TREE USED TO CALCULATE THE JOHNSON ET AL. (2016) NOVELTY METRIC.

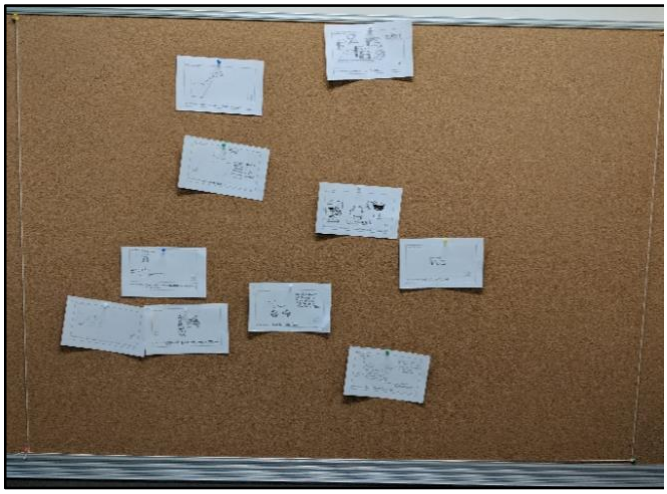

FIGURE 3: EXAMPLE OF HOW EXPERT 2 PINNED ON THE 10 IDEAS ON THE BOARD BASED ON HOW "SIMILAR" THEY WERE TO ONE ANOTHER

physical principle, working principle, or embodiment (see Figure 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 for to power the product (i.e. manual, battery). On the other hand, the working principle was determined by what type of motion was used by the product (i.e. stirring, shaking) and the embodiment was determined by what the product looked like (i.e. shake weight, handheld frother). Total novelty scores were determined by the equations described in [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.

4.2 Qualitative Study

In order to answer our third research question, a second study was conducted with four expert raters to understand what factors they were using to evaluate similarities or differences in design concepts. Specifically, a concept-mapping exercise was used as a direct method for identifying how pairs of design ideas are related to each other and for identifying what design attributes were important in deciding which items were considered novel by raters. The concept-mapping exercise was chosen over other methods such as an interview as it is a more direct method for achieving these goals, due to large number of possible combinations of ideas and attributes involved [70]. While this concept-map may look similar to affinity diagrams [70], an inductive method where ideas are broken up into small chunks and then organized into groups of related information that highlight particular themes, the concept-mapping exercise differed in one major way. Specifically, in affinity maps, the relative position of groups is not meaningful, while in the concept-maps, participants are asked to consider the relative similarity of an idea with all other ideas while placing it. This required moving around entire groups to accurately position them relative to others.

Specifically, for the concept-mapping exercise, four raters were selected from a previous study conducted by Ahmed et al. [71], which asked 11 raters to evaluate 10 milk frother ideas in a survey, where they were provided with 360 triplet queries (all possible permutations of three sketches) and the participants had to decide whether Idea A was more similar to Idea B or Idea C (see details in Ahmed et al. [71]). This set of design sketches was randomly sampled from the larger dataset previously discussed. The internal consistency and cross-rater alignment were computed across all 11 participants and, based on this analysis, four raters were selected who showed high internal consistency and cross-rater alignment. This included one professor (Industrial Engineering), one post-doctoral scholar (Industrial Engineering), one Ph.D. student (Industrial Engineering), and one undergraduate student (Psychology).

These four raters were then asked to complete a second phase of ratings where each participant was provided with the same 10 idea sketches utilized in the triplet survey, printed on 8.5" x 5.5" sheets of paper, see Figure 1 for example sketches. The order of the ideas was randomized for each participant. The raters 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 they were to each other", see Figure 3 for example. The participants were instructed that the sketches were allowed to overlap and the participants were allowed to move the sketches multiple times, until they were satisfied with the idea map created. In addition, the participants were asked to think aloud and the speeches were recorded using video and audio equipment. The audio files averaged 16 mins 48 secs (Standard deviation of 3 mins 40 secs) between the four participants.

The audio was transcribed using NVivo online transcription services [72] and errors from the automatic transcriptions were manually corrected. Figure 3 shows examples of how the sketches were pinned on a board by one of the experts. Importantly, our previous work by Ahmed et al. [71], compared the maps created through this process to the maps created from the triplet survey. The current work, however, shifts the focus of the analysis to the decision-making process of the raters involved in creating these maps.

In order to do this, the audio was qualitatively analyzed sentence-by-sentence using abductive content analysis [73] in NVivo [74]. Abductive content analyses was selected because it has been found to be beneficial in cases studying data with an existing theory - in this case, the novelty-tree developed by SVS [3]) – while also taking into account the variance of data that can be obtained by participants in similar studies [75, 76]. Thus, the analysis of this data started by considering prior literature while also being responsive to the inherent characteristics of the data. In order to do this, open coding was first performed in NVIVO and then through axial coding at intersections where the participant shifted the discussion between ideas. Similar categories were grouped with the intent to understand themes and thought processes of the participants. The categories and sub-categories were directed by the content of the think-aloud recordings and prior research conducted on the same dataset of ideas [32]. The individual nodes were coded under each level of abstraction, particularly the physical principles, working principles and embodiment, as guided by the genealogical tree method proposed by SVS [3]. Comparing the genealogical trees used in the SVS [3] novelty metric, an analogy was assumed as to what constitutes to the physical principle, working principle and embodiment. As most of the sketches failed to dig deep enough into nitty-gritty, the detail level was ignored as suggested by the metric [3]. Two coders independently coded the data and achieved relevant inter-rater agreement to be considered for the analysis. The two raters had a high inter-rater agreement in the analysis process (Cohen's Kappa = 0.88) according to Landis's classification of Kappa [77].

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, a significance level of 0.05 was used in all analyses and effect sizes were classified according to Cohen [78].

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 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 ($r = -.11$, $p = 0.002$). 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 measuring 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 quasi-experts) provided similar values, also known as interrater reliability, for ratings made using the CAT. Using an intraclass correlation coefficient

TABLE 2: CORRELATIONS AMONG CREATIVITY OUTCOMES. ALL BOLD CORRELATIONS STATISTICALLY SIGNIFICANT AT *P* < .05, N = 932; [ICC2 VALUES (I.E., INTERRATER RELIABILITY) VALUES IN BRACKETS].

	Expert Novelty	Expert Quality	SVS novelty	SVS Quality	Quasi expert Novelty	Quasi Expert Quality	Johnson *et al.* novelty
Expert novelty	**[.71]**						
Expert quality	**-.29**	**[.75]**					
SVS novelty	**-.10**	**.30**	**[.85]**				
SVS quality	**-.22**	**.31**	**.17**	**[.62]**			
Quasi-expert novelty	**.74**	**-.34**	**-.11**	**-.30**	**[.78]**		
Quasi-expert quality	**-.41**	**.50**	**.35**	**.32**	**-.50**	**[.56]**	
Johnson *et al.* novelty	**.14**	**.09**	**.39**	**.17**	**.09**	.06	**[.85]**

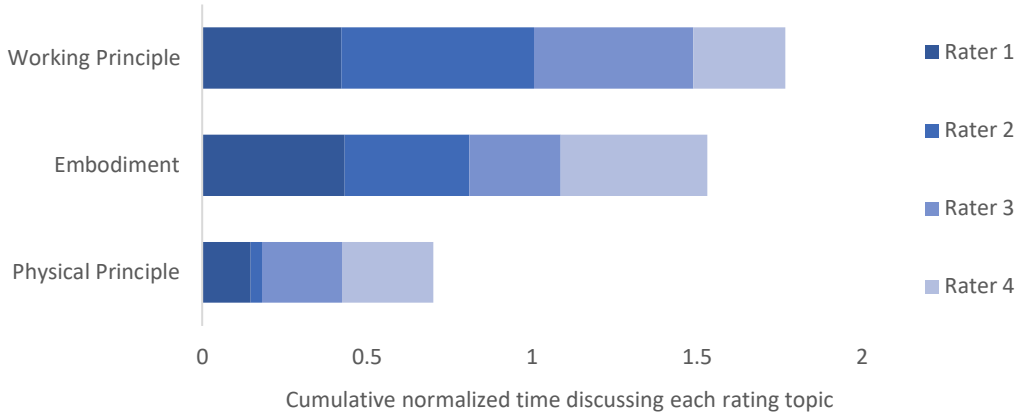

FIGURE 4: CUMULATIVE NORMALIZED TIME SPENT DISCUSSING EACH RATING TOPIC. THE NORMALIZED TIME IS DEPICTED IN ORDER TO ACCOUNT FOR INDIVIDUALS WHO SPOKE MORE OR LESS DURING THE RATING PROCESS.

(ICC2), we found that experts provided similar ratings to one another as depicted by meeting the threshold of .70 [79, 80] for both novelty (ICC2 = 0.71) and quality (ICC2 = 0.75) assessments. While quasi-experts 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 .70 ICC2 threshold (ICC2 = .56). Consistent with trends on the interrater reliability findings, correlations between the aggregated ratings of experts and aggregated ratings of quasi-experts were higher for assessments of novelty ($r = .74$, $p < 0.001$) than quality ($r = .50$, $p < 0.001$), see Table 2.

On the whole, these results suggest that although experts and quasi-experts are capable of providing consistent ratings of novelty, they are less consistent when assessing quality. Although we cannot directly test accuracy 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 quasi-expert 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 quasi-experts 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: What factors do human raters use to evaluate design novelty? How does this compare to the traditional factors used in engineering design research?

The first two research questions identified that while both experts and quasi-experts are capable of providing consistent ratings of novelty, these ratings do not align with traditional novelty metrics used in engineering design research. The question then becomes, why? In order to *begin* to understand what may be causing discrepancies between the CAT and SVS metrics of design novelty, we turned our focus to the data gathered from the qualitative study on human ratings and the subsequent content analysis.

From this content analysis, three main topics and 19 sub-topics were identified, see Figure 4 for the cumulative normalized time raters spent discussing these topics. The factor that was most frequently discussed during the activity was the ***working principle*** of the design (f = 90), which related to the method of frothing. This factor included air (f = 22), spinning (f = 21), movement (f = 19), vibration (f = 10), rotation (f = 6), agitation (f = 5), stirring (f = 5), and the use of a turbine (f = 2). For example, Rater 1 said "Number 5 needs to be close to 2 because it's pressurized air." Along the same lines, Rater 2 said "... idea 8 [has] a similar motion to idea 4 but it's farther away from that bicycle motion."

The second most frequently discussed factor during the qualitative study was the embodiment of the design (f = 76), which related to the ***physical appearance*** of the idea. This factor included containers (f = 23), bicycles (f = 16), beaters (f = 11), pedals (f = 11), shafts (f = 6), centrifuges (f = 4), mixers (f = 3), and pegs (f = 2). For example, Rater 4 said "So right off the bat to me ideas 3, 4, and 7 seem very similar just because they have some pedaling a bike or using a foot pedal in order to get the whole system started." Similarly, Rater 2 said "…Idea number 7. It's pretty similar to idea number 3 because they both have these pedals and they connect to a frother."

Finally, the third most frequently mentioned factor was the ***physical principle*** of the designs (f = 34), which related to the type of power used in the ideas. This factor included human-powered (f = 17), electricity (f = 12) and electrical-power (f = 5). For example, Rater 4 said "I'm going to move idea number 3 closer to [the other] human-power-sourced ideas."

TABLE 3: COMPARISON IN WEIGHT GIVEN TO EACH CATEGORY BETWEEN SVS AND CONTENT ANALYSIS METHODS.

SVS Levels	Content Analysis Themes	SVS weights	Content analysis Weights
Physical Principle	Power source used	10	0.96
Working Principle	Method of frothing	6	10
Embodiment	Form	3	8.66

In order to understand the relative importance of these factors between the SVS and human ratings methods, we scaled the human raters cumulative normalized time spent discussing each topic to a 10-point scale and compared these weights to those assigned by the SVS method (which is out of a 10-point scale). As Table 3 demonstrates, there are large discrepancies in the relative weights assigned to these discussion topics between the two methods. In other words, the qualitative study revealed that humans are using different criteria to evaluate the similarity of design ideas which may contribute to the inconsistencies we see in the novelty scores being assigned by these methods.

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 significant agreement between quasi-expert and expert CAT novelty ratings, there was no significant agreement between these raters for design quality.
- Content analysis revealed significantly different emphasis on what experts and SVS placed on the importance of design features, which may lead to discrepancies in novelty calculations between these methods.

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 for example [15-19] [3, 20-30] [31, 32] [30] [30, 33-37]). The results of our qualitative analysis provide some insights as to why a negative relationship might exist between SVS and CAT expert novelty ratings. Specifically, there was an inequality in what emphasis was placed on facets of the designs when making judgements on design novelty. For example, in the case study presented here, SVS placed higher weight on the physical principle of the designs whereas experts placed higher weight on the working principle. In addition, experts placed almost as much weight on embodiment as they did on working principle, while SVS puts the least amount of weight on embodiment. This may lead to inconsistencies in the ratings provided. Another potential sources 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 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 uniqueness, while SVS uses the term "novelty".

The results in this paper also identified significant agreement between expert and quasi-expert 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 1,000 design ideas. Thus, the results support the use of quasi-experts for novelty assessments in engineering design research when a modified Q-sort technique (Redmond, Mumford & Teach, 1993 [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 quasi-expert 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" (pg. 85).

7. 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 [81]. 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, 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 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" [82].We recommend the following steps as guidelines for promoting a useful nomological network. First, measure known antecedents or contextual predictors of creativity such as autonomy, resource availability, and climate [83]. Second, measure individual differences also associated with creativity, including personality, expertise, and intelligence. Third, 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. Finally, include measures to that provide discriminant validity or evidence that the measure (i.e., CAT or SVS) are 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.

8. LIMITATIONS AND FUTURE WORK

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, future work is needed to compare and contrast these results across 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 difference between expert and quasi-expert 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 quasi-experts in published articles (see for example [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). In addition, while our qualitative data provided some insights on why differences are occurring between these metrics, this study could be further supported by comparing experts and novices in this concept-mapping exercise across a wider range of problems. The weights identified as part of the qualitative study may also have been a function of the number of factors sourced for a given topic and thus should be further investigated.

Finally, creativity has several disputed definitions based on the domain, the environment and processes involved. Our study was an example as to how were 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. Our future work will concentrate 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, that are currently debated across different domains. In addition, we will focus our efforts on identifying how the results present here scale to more complex solutions.

7. ACKNOWLEDGEMENTS

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

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

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

[4] Christiaans, H. H., 2002, "Creativity as a design criterion," Communication Research Journal, 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, Design & Communication in 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," International Journal of Design Creativity and Innovation, 1(3), pp. 131-159.

[7] Johnson, T. A., Caldwell, B. W., Cheeley, A., and Green, M. G., "Comparison and Extension of Novelty Metrics for Problem-Solving Tasks," Proc. ASME 2016 International Design Engineering Technical Conferences & Computers

and Information Engineering Conference, 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 ConferencesChicago, IL, pp. 85-93.
[9] Nelson, B., and Yen, J., 2009, "Refined metrics for measuring ideation effectiveness," Design 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," Proc. IDMME Virtual Concept 2010.
[11] Baer, J., 2012, "Domain specificity and the limits of creativity theory," The Journal of Creative Behavior, 46(1), pp. 16-29.
[12] Casakin, H., and Kreitler, S., 2005, "The nature of creativity in design," Studying Designers, 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," Design Principles & Practice: An International Journal, 5(1).
[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 ConferencesBuffalo, 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," Design Studies, 46, pp. 6-37.
[16] Nikander, J. B., Liikkanen, L. A., and Laakso, M., 2014, "The preference effect in design concept evaluation," Design 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, Springer, 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," Psychology of Aesthetics, Creativity, and the 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 Research, 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," Design 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," Design 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?," Design Studies, 36, pp. 31-58.
[23] Baer, J., 2015, "The Importance of Domain-Specific Expertise in Creativity," Roeper Review, 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 Research Journal, 27(1), pp. 24-30.
[25] Vandevenne, D., Pieters, T., and Duflou, J., 2016, "Enhancing novelty with knowledge-based support for Biologically-Inspired Design," Design 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," Design 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," Design 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," Design 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," Design 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," Design 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," Design 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.
[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," Design 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," Computers & Industrial Engineering, 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," Design 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 and 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," Frontiers in psychology, 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," Journal of Mechanical Design, 142(9).
[39] Amabile, T., 1982, "Social psychology of creativity: A consensusual assessment technique," Journal of Personality and Social Psychology, 43, pp. 997-1013.
[40] Amabile, T., 1983, "Brilliant but cruel: perceptions of negative evaluators," Journal of Experimental Psychology, 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.," Journal of Mechanical Design, 138(2)(2), pp. 021101-021101- 021101-021111.
[42] Sarkar, P., and Chakrabarti, A., 2011, "Assessing design creativity," Design 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 research journal, 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 Research Journal, 22(2), pp. 200-205.
[45] Hennessey, B. A., Amabile, T. M., and Mueller, J. S., 2010, "Consensual Assessment," Encyclopedia of Creativity.
[46] Cropley, D. H., Kaufman, J. C., and Cropley, A. J., 2011, "Measuring creativity for innovation management," Journal of technology management & 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 and Ergonomics in Manufacturing & Service Industries, 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?," International Journal of Design Creativity and 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 Research Journal, 11(4), p. 333/346.
[50] Kaufman, J. C., and Baer, J., 2012, "Beyond new and appropriate: Who decides what is creative?," Creativity Research Journal, 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 Research Journal, 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 Research Journal, 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?," Psychology of Aesthetics, Creativity, and the 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," Organizational behavior and 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 Research Journal, 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 Research Journal, 25(1), pp. 59-74.
[57] Daly, S. R., Seifert, C. M., Yilmaz, S., and Gonzalez, R., 2016, "Comparing Ideation Techniques for Beginning Designers," Journal of Mechanical Design, 138(10), p. 101108.
[58] Shah, J. J., Smith, S. M., and Vargas-Hernandez, N., 2003, "Metrics for measuring ideation effectiveness," Design studies, 24(2), pp. 111-134.
[59] Ter Harr, S., Clausling, D., and Eppinger, S., 1993, "Integration of Quality Function Deployment in the Design Structure Matrix," Cambridge, MA.
[60] Pugh, S., 1991, Total Design, Addison-Wesley.
[61] 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," Journal of Mechanical Design, 133.
[62] Kurtoglu, T., Campbell, M. I., and Linsey, J. S., 2009, "An experimental study on the effects of a computational design tool on concept generation," Design Studies, 30(6), pp. 676-703.
[63] 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," Research in Engineering Design, 24, pp. 65-92.
[64] Besemer, S. P., 1998, "Creative Product Analysis Matrix: Testing the Model Structure and a Comparison Among Products--Three Novel Chairs," Creativity Research Journal, 11(4), pp. 333-346.
[65] Besemer, S. P., and O'Quin, K., 1999, "Confirming the three-factor creative product analysis matrix model in an American sample," Creativity Research Journal, 12(4), pp. 287-296.
[66] 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 ConferencesBuffalo, NY.

[67] Toh, C. A., and Miller, S. R., 2014, "The Impact of Example Modality and Physical Interactions on Design Creativity," Journal of Mechanical Design, 136(9).
[68] 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," Research in Engineering Design, 27(3), pp. 195-219.
[69] 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," Journal of Mechanical Design, 031008: 1-15.
[70] Kawakita, J., 1991, "The original KJ method," Tokyo: Kawakita Research Institute, 5.
[71] Ahmed, F., Ramachandran, S. K., Fuge, M., Hunter, S., and Miller, S., 2019, "Interpreting Idea Maps: Pairwise comparisons reveal what makes ideas novel," Journal of Mechanical Design, 141(2), p. 021102.
[72] QSR, 2018, "NVivo Transcription services," https://www.qsrinternational.com/nvivo/nvivo-products/transcription.
[73] Timmermans, S., and Tavory, I., 2012, "Theory construction in qualitative research: From grounded theory to abductive analysis," Sociological theory, 30(3), pp. 167-186.
[74] Software, N. Q. D. A., 2012, "QSR International Pty Ltd. Version 10."
[75] Charmaz, K., and Belgrave, L., 2012, "Qualitative interviewing and grounded theory analysis," The SAGE handbook of interview research: The complexity of the craft, 2, pp. 347-365.
[76] Creswell, J. W., and Miller, D. L., 2000, "Determining validity in qualitative inquiry," Theory into practice, 39(3), pp. 124-130.
[77] Landis, J. R., and Koch, G. G., 1977, "The measurement of observer agreement for categorical data," biometrics, pp. 159-174.
[78] Cohen, J., 1988, "The t test for means. Statistical power analysis for the behavioural sciences," Hillsdale, NJ: Lawrence Erlbaum Associates.
[79] James, L. R., Demaree, R. G., and Wolf, G., 1984, "Estimating within-group interrater reliability with and without response bias," Journal of applied psychology, 69(1), p. 85.
[80] Lance, C. E., Butts, M. M., and Michels, L. C., 2006, "The sources of four commonly reported cutoff criteria: What did they really say?," Organizational research methods, 9(2), pp. 202-220.
[81] Case, B., 1994, "Walking around the elephant: A critical-thinking strategy for decision making," The Journal of Continuing Education in Nursing, 25(3), pp. 101-109.
[82] Cronbach, L. J., and Meehl, P. E., 1955, "Construct validity in psychological tests," Psychological bulletin, 52(4), p. 281.
[83] Ma, H.-H., 2009, "The effect size of variables associated with creativity: A meta-analysis," Creativity Research Journal, 21(1), pp. 30-42.