But is it creative? Delineating the Impact of Expertise and Concept Ratings on Creative Concept Selection

Christopher A. Gosnell, first author
The Pennsylvania State University
112 Leonhard Building,
University Park, PA 16802
cag5266@psu.edu
ASME Membership (if applicable)

Scarlett R. Miller, second author [1]
The Pennsylvania State University
213-P Hammond Building
University Park, PA 16802-1401
scarlettmiller@psu.edu
ASME Membership (if applicable)

Keywords: conceptual design, design evaluation, design theory and methodology, product design

ABSTRACT

While creativity is often stressed in the conceptual phases of design, it is rarely considered during the concept selection process. Before effective methods can be developed to aid in creative concept section, however, differences in perceptions of creativity between expert and novice designers and the influence of creativity evaluation methods on the process must be considered. Therefore, this paper was developed to address these questions by studying 11 expert and 11 novice designers. Specifically the study was developed to understand if experts' and novices' perception of a concepts creativity aligned, to introduce and compare the utility of our Tool for Assessing Semantic Creativity (TASC) to existing creativity evaluation methods, and to identify if

[1] Corresponding author information can be added as a footnote.

our TASC method could be used as a proxy for expert evaluators. Our findings reveal that expert and novices generally had similar perceptions of a concept's creativity and that the TASC method was tapping into similar constructs of human perceptions of concept creativity. The results of this study contributes to our understanding of the factors that influence the selection/ filtering of creative ideas *after* idea generation and provides a framework for research in this field.

1.0 INTRODUCTION

Innovation is a crucial component of long-term economic success [1]. As such, engineering design research has long since devoted attention and resources to developing tools and methods for supporting creativity during idea generation (see for example [2-5]). While the goal of these methods is to help designers generate a large quantity of effective solutions and explore a larger solution space [6], the creative ideas developed through these methods are often rapidly filtered out during the concept selection process [7]. In other words, while creativity is often emphasized in the early phases of design, it is rarely emphasized in the later stages [8]. This is problematic because even if designers develop creative concepts, these concepts may not be selected to move forward in the design process. In fact, many companies have acknowledged that they often perform poorly at selecting their own most promising ideas [7], which may hinder the innovation potential of companies. While selecting creative concepts is a vital component of the design process, few tools exist for helping designers quickly and accurately judge the creativity of design ideas during the concept selection process [9].

While not specifically focused on creativity, there has been a wealth of research devoted to developing methods for aiding designers in decision-making during the concept selection process. Broadly, these methods fall into five major categories: Utility Theory [10-12], Analytic Hierarchy Process (AHP) method [13-15], Pugh's evaluation method [16-18], Quality Function Deployment (QFD) matrix method [19, 20], and fuzzy-set methods [21, 22] (see [23] for discussion). While these methods are widely used in academic and industrial practices for evaluating concepts, they often neglect to consider the creativity or uniqueness of each concept during the selection process [24].

While recent studies have begun to explore new concept evaluation methods that focus on both the quality and novelty of the design ideas developed during concept selection (see for example [9, 25]), these methods are largely unexplored for their impact on creative concept selection or their ability to aid decision makers in the process. In addition, while there have been metrics and methodologies developed to help designers evaluate engineering design concept creativity [3, 26-28], these methods are rarely used outside of academic purposes due to the time and in-depth process required to analyze each design concept. Therefore, new methods are needed for properly evaluating design concept creativity in order to help designers more thoughtfully consider creative concepts during the concept selection process.

A less time-intensive, qualitative approach for evaluating concept creativity is to rely on independent reviewers' subjective agreement [29]. This method is based on the consensual definition of creativity that states that an idea is creative if a group of independent reviewers subjectively agree that it is creative. While this method provides a more efficient means of evaluating concept creativity, the quality of these judgments relies on the evaluators' knowledge and expertise in the subject domain [30]. Despite the speed behind human perception, however, judgments can be inconsistent and lack quantitative support [31, 32]. In addition, while expert designers are often used to evaluate candidate designs based on their experience, there has been little research geared at exploring the difference between expert and novice ratings of concept creativity.

Therefore, the purpose of this paper is two-fold. First, we seek to identify perceptual differences in concept novelty and quality [29, 33] between expert and novice engineering designers across three different problem domains. Second, we seek to introduce and test a novel method for evaluating the absolute creativity (both the novelty and quality) of design concepts using adjective selections and semantic similarity. This approach minimizes human biases and the costs (time and money) required for finding, meeting, and training skilled raters. This work contributes to our understanding of the utility of new metrics for evaluating creativity and directs us to a more efficient system for evaluating design concepts during concept selection.

2.0 Methods for Evaluating Design-Concept Creativity

A significant amount of research has been directed towards understanding how designers make decisions during concept selection in order to develop tools to improve decision-making. For example, in engineering design research has led to the development of metrics to determine the effectiveness of concept generation sessions with respect to creativity [27]. The majority of this research has focused on relative measures of a concept's creativity compared against other ideas in the same generated set [34, 35]. The relative nature of these measures help to inform the designers about the uniqueness of the ideas *with respect to a specific design problem within a set of ideas developed* [36, 37]. In this way, designs generated in the same design session addressing the same problem can be compared and contrasted to tease out designs to develop further.

In the field of engineering design, relative creativity is often measured by breaking down the design concepts into their unique features [38]. For example, the widely adopted Shah, Vargas-Hernandez, and Smith (SVS) method computes overall design novelty based on "how unusual or unexpected an idea is compared to other ideas. Not every new idea is novel since it may be considered usual or expected to some degree"(pg. 117) [6]. Through this process of decomposition, researchers are able to compare and contrast each individual design using feature-tree analysis such as the comparison of a designs shape, color or purpose [36, 37]. 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 occur more frequently in the design set. This method of decomposition and feature-tree analysis has become a gold standard in engineering design research due to the limited rater bias and repeatability [6, 27]. Despite the wide use of this method, however, many limitations have been reported such as the extensive training needed to combat low inter-rater reliability and the difficulties interpreting multiple SVS metrics simultaneously [39, 40].

Because of these challenges, cognitive scientists have adopted a vastly different approach for evaluating concept creativity by subjectively evaluating design concepts based on a design's quality (functional ability), originality, elegance and the variety of concepts generated [41]. This evaluation begins with the selection of anchor concepts from the idea set that represent high, medium, and low creativity [42]. With these anchors, judges are *trained* to evaluate other concepts on a relative basis (how creative are they compared to the set of ideas). Afterward, the concepts generated are evaluated using 7-point Likert scales. This method has been used widely to assess creativity and has been praised for its strong inter-rater reliability values in the range of 0.80–0.90 [43, 44]. Despite the widespread adoption of this method in cognitive science, this method requires careful selection of the anchoring design examples and extensive training of the rating team [1, 45].

Because of the deficits of existing approaches, researchers have begun to explore alternative methods for evaluating concept creativity through the development of Computation Design Creativity systems (CDC) [46]. CDCs provide an opportunity to leverage computational power and review large data sets and potentially measure an ideas historical creativity, or the fundamentally novelty of an idea with respect to the whole of human history [47]. This is important because the development of more robust creativity frameworks could be the key to enabling CDC systems. For example, the work of Maher, and Fischer [48] has sought to more appropriately characterize product creativity for use within CDC systems using the characteristics of novelty, value and surprise. In addition, the work of Gero and Kannengiesser [49, 50] has also sought to enhance CDC systems through the development of an ontological framework using the creativity characteristics of the designs function, behavior and structure. Their proposed system enables the identification of creativity within the product and the process by looking at the interactions between the expected, interpreted and external worlds of these characteristics. Although computational power is readily available, it has been challenging to adopt more recognized creativity metrics, such as the SVS method, into a computer based system [6, 46].

The deficit of current evaluation methods and the emergence of CDC systems provide an opportunity for new creativity evaluation metrics. Therefore, the goal of this research is test the effectiveness of a new, global creativity evaluation method and compare this approach to human perceptions of in engineering product design.

2.1 Cognitive Evaluation of Design Creativity and the Role of Experience.

Because of the variability of human judgment in the design process, it is important to understand the influence of experience and biases in concept evaluation. Cognitive psychology research has shown that expertise is linked to the development of automatic processing of relevant information due to pattern recognition [51-53]. It would follow then that, when solving problems or designing, it may be possible for experienced designers to make reasonable decisions based on automatic processing, which allows information to be quickly sorted and used. However, this automated processing may also lead to individuals disregarding important, or subtle information that an inexperienced individual will retain [54].

While there is a general support of the use of expert raters in the cognition literature, it was only recently that engineering design researchers began to explore the role of expertise on design-concept ratings. Specifically, a recent study by Green et al. [55] showed that it may be possible to use novice designers to evaluate design creativity with minimal training while still achieving expert-level feedback. In this study, cumulative students ratings were compared to an expert rater. The results showed a high inter-rater agreement between student and expert ratings of design concepts. However, the finding was found with a minimum of 40 novice student raters and was only tested with one design task, which limits its utility in practical settings like engineering education and industry.

When exploring the impact of expert versus novice raters it is also important to consider the problem-solving and decision-making strategies that guide experienced and inexperienced designers [56, 57]. For example, a recent case study of an experienced industrial designer showed that small heuristics are often used to effectively explore the problem space and develop more creative solutions [5]. While not explicitly explored in this study, these smaller, and quickly formulated decisions by experienced designers might also impact the concept-selection phase in the design process. In another study, experts were shown to describe concepts more efficiently and produce sketches that contained less detail than non-experts [58]. These findings align with prior cognition research regarding automatic processing of information due to expertise and context [59].

While current research has outlined that experts and novices may view creativity differently due to prior experience and decision making heuristics, there is still a limited understanding of how designers perceive and evaluate early phase design ideas. Without this knowledge, it is impossible to develop new methods or tools to support the evaluation of candidate concepts. Therefore, the current study was developed to understand the success and limitations of current creativity evaluation methods for mimicking expert opinion. In this way, improvements and modifications can be made to strengthen the capabilities of future evaluation tools.

2.2 Affective Engineering Techniques

Understanding the subjective nature of human needs has been key to the development of Affective, or Kansei, engineering practices that seek to use consumer affective needs to design products [60]. Affective design refers to the process of creatively engaging the customer’s emotions to differentiate one design from another [61]. In order to achieve this, researchers have utilized Kansei methods to quickly evaluate human perception [62], satisfaction [61] and desirability [26] as a means to develop innovative product designs.

Kansei engineering generally includes the identification of the design problem, generation of design samples, sharing the samples with potential customers, and finally, analyzing the adjectives used by the customers to describe the design samples [63]. This process of obtaining adjectives helps designers to create a model for how customers interpret the designs. Specifically, during this method, adjectives are clustered by how well they match a specific design factor [64]. This categorization process is similar to that of the relative measures of creativity involving feature-level analysis [40] but, instead of comparing and contrasting unique features, it uses adjective comparisons that are not limited by the design space being explored. The clusters of words are generally formed by the emotional response that can be elicited by adjectives such as “fresh”, “genuine” or “appealing” [65]. Based on these clusters, contrasting words are then collected and 7-point Likert scales with bi-polar adjectives on each end are established. An example of adjective pairs could be “hot–cold”, “unique–conventional” or “feasible–impossible”. With the sets of words defined, perceptions about different design features, concepts, or full products can be obtained by surveying a panel of customers using these polarized Likert scales and performing multivariate analysis [61, 64, 65]. This method of design analysis has the rigor and relevance of Shah, Vargas-Hernandez, and Smith’s method [66], but embraces the subjective nature of creativity and design.

While Kansei engineering applies relatively strict procedures and statistical analysis to understanding human perception, the work of Benedek and Miner has looked at the perception of design desirability in a more qualitative fashion [26]. Their work has resulted in the development of Product Reaction Cards to help enable the discussion and feedback from participants regarding the desirability and usability of product designs using adjectives on the cards. The method involves presenting a participant with a design(s) and asking them to choose five of the cards that describe how the design(s) make them feel [26]. Participants are then asked to provide feedback on why the words were chosen.

While the Kansei engineering methods do not rely on scales or questionnaires and do not require participants to generate the works on their own, the utility of these methods have not been explored in the concept selection process. Therefore, while potentially useful, empirical studies are needed to explore the use of these methods in an engineering design context. Therefore, the current study was developed to explore the use of affective engineering techniques, and compare this method to existing relative methods.

2.3 Tool for Assessing Semantic Creativity

In order to combat some of the deficits of both human perception and relative creativity metrics and leverage the computational power of CDC's, the authors have developed a new concept evaluation tool called the Tool for Assessing Semantic Creativity (TASC). This tool was developed to create a global evaluation of a concepts creativity in order to provide explore the fundamentally novelty of the ideas with respect to the whole of human history (*historic creativity*) [47]. This is in contrast to other approaches like SVS that measure relative creativity by reducing the concept evaluation space to only include only concepts developed during a single design session or for a single design problem [34, 67]. While our approach is explained in detail in section 3, it is important to highlight how the tool works and the rationale for the tool in order to lay the foundation for our research study.

TASC is based on the foundational work of Benedek and Miner [26] who developed an industrial design decision-system that uses a set of carefully selected words to describe a users reaction to different product concepts. This method requires individuals to select words from a set of adjectives that they feel best describes their feelings towards the design concept. Roughly 40% of the words in the set are considered 'negative' in order to helps evaluators provide more rounded feedback on the concepts and not bias the decision maker. The purpose of this tool was to help participants describe intangible aspects of a products desirability such as 'desire' and 'fun'. Although this system does not generate a quantitative score for the design concepts, in fact the results are usually presented as a list or visualization of the words chosen, it does presents a simple method for obtaining evaluations from decision makers that minimizes the biases associated with asking individuals to merely 'evaluate a concept'.

Like Benedek and Miner's Toolkit, TASC requires participants to select adjectives to describe the idea and then uses natural language processing and Latent Semantic Analysis to develop a creativity score for each idea. This type of word analysis has been instrumental in applications such as search engine optimization [68], consumer specific marketing tools [69] and data mining [70] which lend themselves to extracting value and making decisions from natural language autonomously. The rational for this approach in concept evaluations it that by combining semantic evaluations with a word selection task we may be able to minimize biases associated with pure human judgments while maintaining some of the consistency and reliability of using a more quantitative approach. The idea for this method is supported by other work on creative word selection that has shown that semantic similarities between words can be used to measure participant creativity [71]. While TASC has the potential to aid in creative concept selection, no study to date has explored its effectiveness. Therefore, the goal of this study is to develop and test this method by comparing it to existing concept creativity metrics.

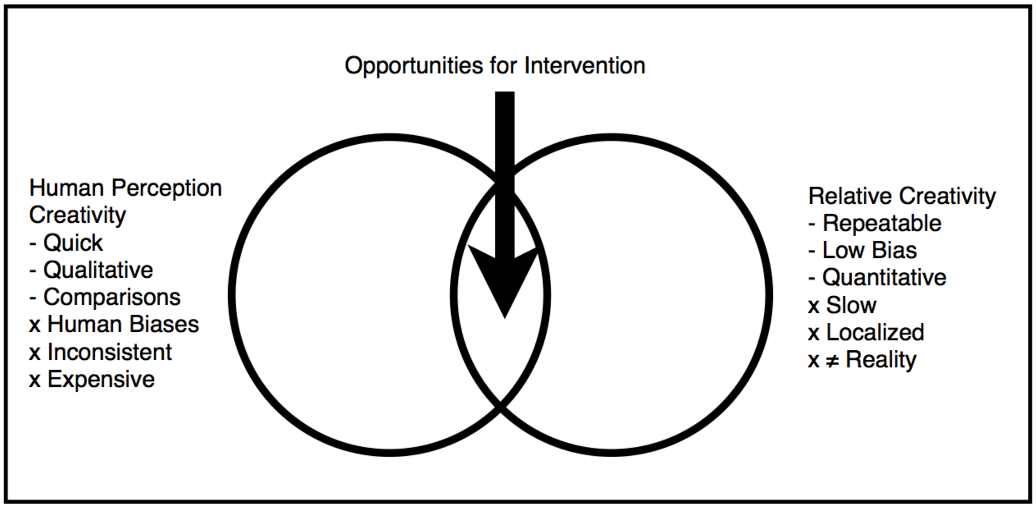


Figure 1 Venn diagram comparing design creativity evaluation methods

3.0 Research Objectives

Prior work has discussed the role of experience in concept evaluation as well as the many tools used in engineering design to evaluate design creativity. However, as the prior literature brought to light, there are opportunities for interventions that utilize both the repeatability and quantitative nature of creativity metric and the efficiency of human perception, see Figure 1. Specifically, this image illustrates that while human perception is quick and thus efficient, it is subject to cognitive biases and limitations that can lead to inconsistent reviews. On the other hand, the relative measures that have been developed, while repeatable, are time intensive to develop limiting their utility. Therefore, it is important that we understand the impact of methods that rely on the efficiency of human perception and have the repeatability of the more standardized methods.

The introduction of new concept evaluation methods may serve to overcome some of the barriers of both relative and global creativity metrics. In order to test this theory, in this paper we introduce and test a novel method for rating design concept creativity called Tool for Assessing Semantic Creativity (TASC) that relies on the calculation of a creativity score based on adjective selections. While this method may prove useful for concept evaluation, it has yet to be explored.

Therefore, the purpose of this research is to understand the impact of rater experience on creativity assessment methods, and how this knowledge can be used to improve concept selection tools. Specifically, our study was developed to answer the following questions:

1. What are the similarities and differences between experts' and novices' perceptions of designs novelty, quality and overall creativity? Prior research in cognitive science has identified that novices can become easily distracted by a design's relative newness and focus heavily on this aspect of creativity [72, 73]. In addition, research has shown that novices tend to rely heavily on personal experience to evaluate design feasibility which may lead to

inaccurate perception because they lack the personal and domain experience of experts [74, 75]. Therefore, we hypothesize that there will be differences among expert and novice perceptions of early phase ideas in each of these areas.

2. How does the Tool for Assessing Semantic Creativity (TASC) that is based on word evaluations (described in the following sections) compare to human perception of creativity and the relative SVS [6] method? Although relative measures of concept creativity allow for reliable and repeatable measures of creativity in engineering design research, they are timely to implement and require substantial training of raters to attain sufficient inter rater reliability [46, 76]. On the other hand, using human perception to rate concept creativity is faster but is limited by the cognitive biases of the decision maker [77, 78]. Therefore, our hypothesis is that our TASC method will tap into constructs of both relative creativity measurements and human perception resulting in a more global assessment of design creativity.
3. Do the TASC and SVS methods align with expert human perception? If so, can TASC be used as a proxy for expert ratings? Prior research has shown contradictory findings on whether or not novices can produce expert-level evaluations. In some literature, novices have been cited as being weaker in their abilities to evaluate design creativity due to their lack of experience [79, 80], whereas more recent literature reports that it is possible to obtain expert-level ratings from 40 trained novices [55]. Because of this, we

hypothesize that our TASC method used by novices will be able to obtain similar evaluations as experts due to the use of adjective selections that are not experience dependent.

4.0 Methodology

To answer these research questions, a controlled study was conducted with a total of 22 engineering design participants. This section summarizes the methodological approach taken to conduct this study.

4.1 Participants

The participants in this study were recruited via email to engineering design list serves. In total, 22 engineering designers (11 females, 11 males) with experience ranging from undergraduate education to 30 years of industry experience were offered $15 as remuneration for participation in this study. Participants with fewer than three years of engineering design experience were considered novices (N=11) while the remaining eleven participants were considered expert engineering designers in the study (N=11). The experts were identified using a two prong classification system; first, the individual had to have a minimum of 3 years of design experience and second, they had to rate themselves at least a 3 on Likert survey question on expertise where a 1 was a novice and a 5 was considered an expert. Three was chosen as the cutoff for expertise in this second prong of our classification code due to central tendency biases on Likert scale survey items. Ten of the eleven experts in our study had engineering design–related advance degrees ranging in areas of focus from human computer interaction to automotive textile product design.

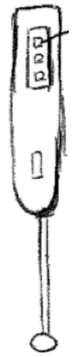
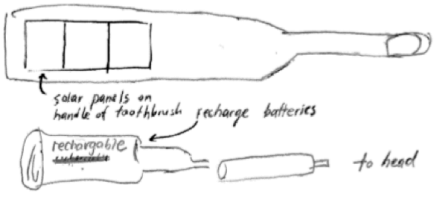
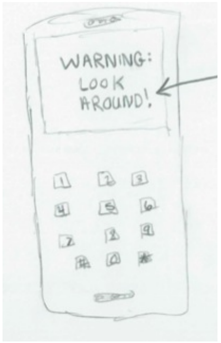
4.2 Experimental Procedure

At the beginning of the study, the procedure and purpose of the study was presented to the participants and any questions were answered. Next, an IRB document was completed along with an 81-question survey where participants were asked to rate 9 design concepts from three different design tasks (27 total designs) using two different methods: (1) an Adjective Selection Questionnaire (ASQ) and (2) a Perceived Creativity Rating Scale, see Figures 2 and 3 for example survey items and procedures. The design concepts and design tasks were randomized for each participant to help control for learning effects. The details of the questionnaire and design concepts rated are provided in the following sections

Design Concepts

The 27 design concepts selected to test our method were taken from three prior research studies conducted by the authors. In these prior studies three design tasks were presented: (1) "Design a novel milk frother" [81], (2) "Design a novel power mechanism for an electric toothbrush" [82], and (3) "Design a device that minimizes accidents on campus from walking, and texting or walking and listening to an MP3 player" [83], see Appendix for the details of these tasks. These design tasks were selected for the current study to represent a range of design problems from well defined (toothbrush problem) to open-ended (walking around campus problem). This was done because current methods have been criticized for their inability to easily be implemented for multiple problem domains [46].

Table 1: Three of the design concepts and their ratings used in the study.

		 solar panels on handle of toothbrush recharge batteries rechargable to head	
Design Problem	**Innovative Milk Frother**	**Toothbrush Power Mechanism**	**Reducing accidents on campus**
SVS Quality Score	high	medium	high
SVS Novelty Score	medium	high	low

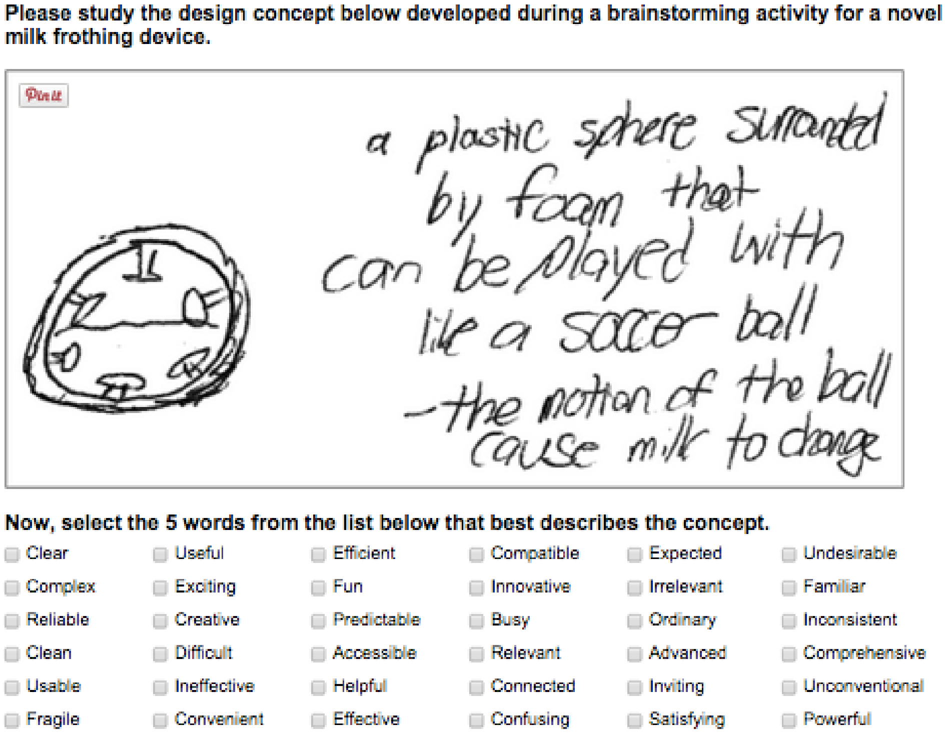
The 27 design concepts selected for this study were analyzed using the SVS method's novelty and quality measures (see [6] for description of this procedure in prior studies) [81, 83, 84]. Of the ideas generated in these prior studies, nine ideas were selected from each of the three design problems in order to represent all combinations of high, medium, and low novelty, and high, medium and low quality (e.g., and idea with high novelty and low quality), see Table 1 and Figure 4 for a demonstration of some of the concepts selected for the study.

Part 1: Adjective Selection Questionnaire (ASQ)

During the first part of the survey (the ASQ, which is the first component of TASC) participants were provided with a brief description of one of the three design tasks and instructions for the rating method. For example, for the milk frothing task participants were provided with the following description,

> "*In this section, you will be presented with design concepts that were developed by engineering students. These students completed a brainstorming task where they were asked to develop concepts for a* ***novel device that froths milk effectively****. In the following questions, you will be presented with design concepts developed for the design task above. You will be asked to select the 5 words you feel best describe the concepts presented. You must select 5 words for each concept."*

Next, participants were asked to "study the design concept below developed during a brainstorming activity for a novel milk frothing device" and then "select the 5 words from the list below that best describes the concept" (see Figure 2 example question). Specifically, the Adjective Selection Questionnaire (ASQ) asked participants to rate each of the 27 design concepts one at a time and select five adjectives from a list of 36 words that best described the concept being evaluated. The 36 adjectives used in the ASQ were derived from the Microsoft Desirability Toolkit (MSDT) which was developed in prior studies to test the utility of word selections for measuring the desirability of design concepts [26, 85]. The MSDT contains a list of 55 words that were selected and tested in three field studies [26]. In the current study, we analyzed these 55 words for their semantic similarity, or relative likeness in meaning [86], to the words innovative and feasible using the software tool DISCO, because design creativity is often described as ideas that are both novel and technically feasible [6, 9].



Please study the design concept below developed during a brainstorming activity for a novel milk frothing device.

a plastic sphere surronded by foam that can be played with like a soccer ball -the motion of the ball cause milk to change

Now, select the 5 words from the list below that best describes the concept.

☐ Clear
☐ Complex
☐ Reliable
☐ Clean
☐ Usable
☐ Fragile
☐ Useful
☐ Exciting
☐ Creative
☐ Difficult
☐ Ineffective
☐ Convenient
☐ Efficient
☐ Fun
☐ Predictable
☐ Accessible
☐ Helpful
☐ Effective
☐ Compatible
☐ Innovative
☐ Busy
☐ Relevant
☐ Connected
☐ Confusing
☐ Expected
☐ Irrelevant
☐ Ordinary
☐ Advanced
☐ Inviting
☐ Satisfying
☐ Undesirable
☐ Familiar
☐ Inconsistent
☐ Comprehensive
☐ Unconventional
☐ Powerful

Figure 2: Example question from the ASQ for the milk frother design problem.

DISCO is an online and downloadable Java class that computes the distributional similarity between words using co-occurrences [87]. For example, although the words “cake” and “eat” have similar occurrences within a text the words “cake” and “pie” are closer in similarity. DISCO looks at these word relationships at multiple levels of contextual relatedness, and similarity of the word’s meanings. We used these calculations of semantic distance to identify words that represented a “60% positive and 40% negative/neutral” relationship to the words innovation and feasibility in an effort to minimize participant selection bias as has been done in prior studies [26]. It should be noted that a negative value was assigned to negative/neutral adjectives during the coding process in order to account for bias [26]. DISCO was used in the current study due to its strong correlation with human judgment [87]. The semantic distances calculated during the selection process were used to create two numeric indices of weights for each adjective, for details on semantic weights please see section 3.0 below. The complete list of 36 words used in the current study, and their respective semantic weight for feasible, and innovative can be seen in Table 2. It should also be noted that, during the study, the order in which the participants saw each problem and each idea within each problem was randomized to reduce ordering effects.

Table 2: Index of the 36 adjectives for evaluators to choose from including TASC semantic weights used for calculations (Innovative weight, Feasibility weight)

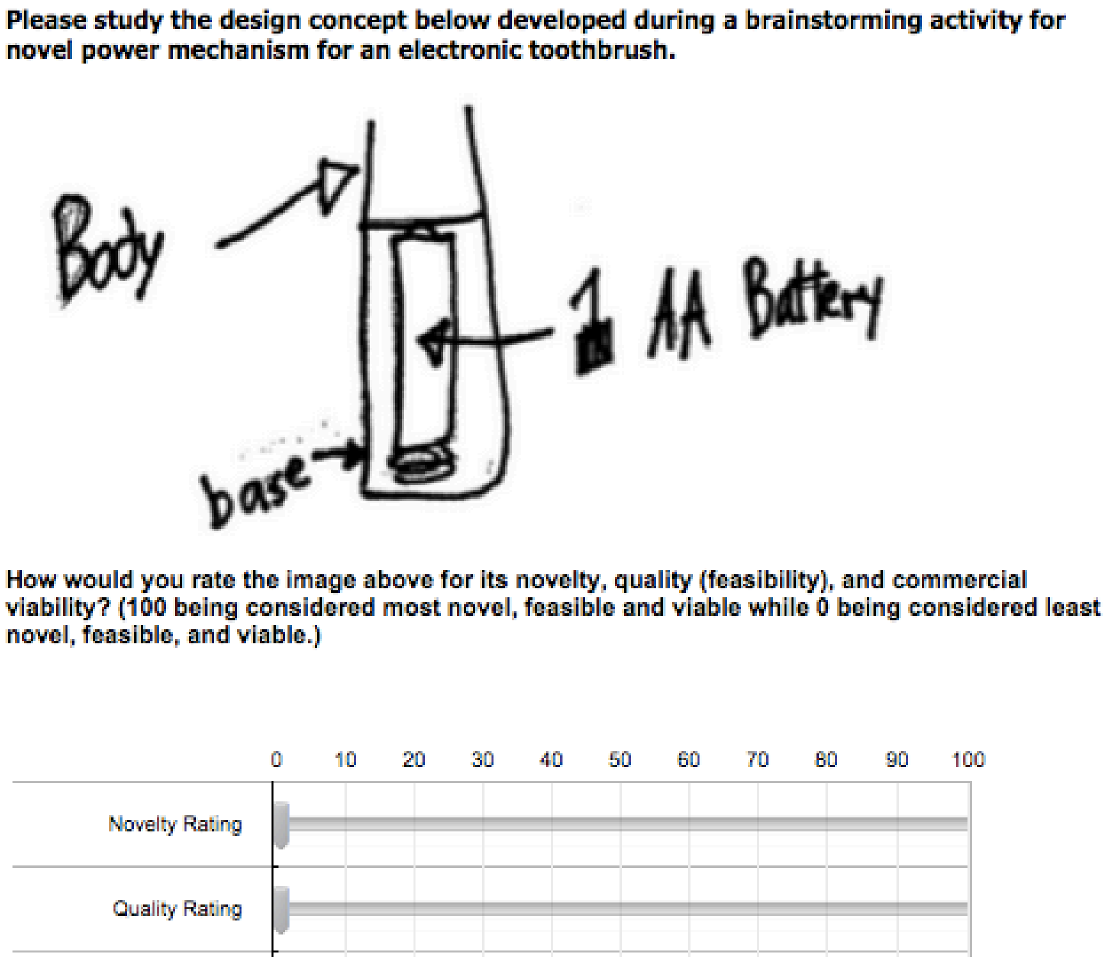
Accessible (0.32,0.39)	Fragile (−0.36,−0.38)
Advanced (0.46,0.30)	Fun (0.21,0.20)
Busy (−0.25,−0.27)	Helpful (0.34,0.41)
Clean (0.29,0.29)	Inconsistent (−0.38,−0.44)
Clear (0.40,0.43)	Ineffective (−0.29,−0.44)
Compatible (0.30,0.26)	Innovative (1,0.36)
Complex (−0.39,−0.32)	Inviting (0.08,0.07)
Comprehensive (0.49,0.29)	Irrelevant (−0.28,−0.46)
Confusing (−0.38,−0.44)	Ordinary (−0.30,−0.26)
Connected (0.13,0.18)	Powerful (0.38,0.31)
Convenient (−0.43,−0.46)	Predictable (−0.40,−0.45)
Creative (0.56,0.32)	Relevant (0.47,0.42)
Difficult (−0.39,−0.51)	Reliable (0.50,0.47)
Effective (0.44,0.43)	Satisfying (0.30,0.37)
Efficient (0.51,0.46)	Unconventional (0.57,0.32)
Exciting (0.43,0.32)	Undesirable (−0.34,−0.36)
Expected (−0.18,−0.29)	Usable (0.33,0.38)
Familiar (−0.45,−0.36)	Useful (0.51,0.49)

Part 2: Perceived Creativity Ratings

Once participants completed the ASQ the participants completed the Perceived Creativity Ratings part of the survey. During this stage, each participant was again presented with a brief description of the design task and then provided instructions on how to rate the ideas. Specifically, for the toothbrush power mechanism design problem, participants saw the following instructions:

> "In this section of the survey, you will again be presented with the design concepts that were developed by engineering students. These students completed a brainstorming task where they were asked to develop concepts for a **novel power mechanism for an electronic toothbrush**. In the following section, you will be asked to rate the previous design concepts on a scale from 0 to 100 regarding the concept's novelty, feasibility and commercial viability. A rating closer to 0 will be considered less novel or feasible while a rating closer to 100 will be considered more novel or feasible. **Definitions for reference: Novel -** *how unusual or unexpected*; **Feasible -** *possible to do easily and how well it meets design specifications;* **Viable -** *able to complete effectively and make a profit"*

In other words, instead of having the participants select adjectives describing the ideas, they were asked to evaluate the concept on a sliding scale from 0–100 for the concept's novelty and feasibility, with 0 being least novel/feasible and 100 being most novel/feasible, see Figure 4. The order in which the participants saw each of the three tasks and each of the 9 design tasks was again randomized. Once the perceived ratings were complete, the study was concluded.



Please study the design concept below developed during a brainstorming activity for novel power mechanism for an electronic toothbrush.

How would you rate the image above for its novelty, quality (feasibility), and commercial viability? (100 being considered most novel, feasible and viable while 0 being considered least novel, feasible, and viable.)

Figure 3: Example question from the perceived creativity ratings portion of the survey.

4.3 Metrics

Once the study was complete, several metrics were created to compare the SVS relative metrics, human perception, and our global TASC method. These metrics are described in detail in the following sections.

Design Novelty

Novelty was calculated in prior studies by the authors using the SVS method [81, 83, 84]. SVS defines novelty to be “how unusual or unexpected an idea is as compared to other ideas” (p. 117) [6]. In this way, SVS-inspired methods generally look at novelty in a relative fashion, where concept novelty is compared to the other ideas developed for a given problem domain. In other words, these types of metrics do not take into account other products on them without taking into account a design’s novelty with respect to all of history [1].

Novelty, in this analysis, was calculated based the novelty of each feature within a design in comparison to the features within all of the designs being reviewed [81]. Ultimately, these calculations produce a value between 0 and 1. Designs with novelty values closer to 0 indicate less novel concepts. Conversely, novelty values closer to 1 indicate concepts that are more novel. The 27 design concepts selected in the current study were selected to represent ideas with low, medium, and high novelty for each of the three design tasks explored (frothing milk, powered toothbrush and safe texting).

Design Quality

The 27 design concepts used in the study were also analyzed using the SVS method for quality (see [81] for in-depth discussion) [88]. They define quality to be “the feasibility of an idea, and how close it comes to meet the design specifications,” (p. 117) [6]. In the current study, the quality values were calculated by having evaluators answer the following questions, “Does it complete the task?”, “Is it technically feasible to execute?” and “Is it technically easy to execute?” By answering these questions, quality is evaluated on a 3-point scale that is normalized (by dividing the human responses by 3) to attain a score between 0, and 1 with 1 considered the maximum absolute quality rating. Once these calculations were complete, the 27 design concepts were selected for the current study to represent ideas with low, medium, and high quality for each of the 3 design tasks explored (frothing milk, powered toothbrush, and safe texting).

Design Creativity

Overall design creativity was calculated as a function of the design novelty and quality scores that utilized the SVS method [6]. Design creativity of the 27 designs was calculated by taking the direct sum of the design novelty and design quality scores from each design. Prior studies have shown how novelty and usefulness parameters can be combined to produce an overall assessment of creativity [1].

With creativity ratings from each participant (evaluator), aggregate perceived creativity ratings can be calculated by averaging participant ratings for each design. These scores were then used to rank the ideas according to their design creativity score by assigning a value of 1 (most creative) to the design with the highest design creativity score, and 9 (least creative) to the concept that had the lowest perceived creativity score. This was completed for the nine designs within each problem domain (milk frother, toothbrush and texting).

TASC Metrics Overview

In addition to SVS, the TASC metric was also calculated from the ASQ. The TASC metric seeks to provide an absolute measure of concept creativity in order to provide an opportunity to evaluate and compare design concepts irrespective of different problem sets. The three TASC scores (innovation, feasibility and creativity) are described in detail in the following sections.

TASC-innovation

TASC-innovation is calculated to provide a global assessment of concept novelty. In order to calculate this, the innovation semantic weights for each of the five words chosen by each participant for each design concept was summed where S_n is the semantic weight of word n, and I_{ijk} is the innovation rating for each i (design concept), j (design problem) and k (evaluator). This calculation results in a value between −1 (meaning low novelty) and 1 (meaning high novelty). The method of computing I_{ijk} is

$$I_{ijk} = \frac{\sum_{n=1}^{5} S_n}{5}. \qquad (3\text{-}1)$$

It should be noted that innovation was used in this methodology instead of the word novelty due to the fact that the semantic system could not distinguish between the word novel (book) and novel (innovative). After completing this for each participant's Adjective Selection Questionnaire (ASQ) response, aggregate TASC-innovation ratings were completed by averaging the ratings from each participant for each design within expert, and novice groups. These scores then were used to rank the ideas according to their TASC-innovation score by assigning a value of 1 (most novel) to the design with the highest TASC-innovation score, and 9 (least novel) to the concept that had the lowest TASC-innovation score. This was completed for the 9 designs within each problem domain (milk frother, toothbrush and texting).

TASC-feasibility

TASC-feasibility is calculated to provide a global assessment of concept feasibility. In order to calculate this, the feasibility semantic weights for each of the five words chosen by each participant for each design concept was summed where S_n is the feasibility semantic weight of word n and F_{ijk} is feasibility rating for design concept i, design problem j, and evaluator k. This calculation results in a value between −1 (meaning low feasibility), and 1 (meaning high feasibility). The method of computing F_{ijk} is

$$F_{ijk} = \frac{\sum_{n=1}^{5} S_n}{5}. \qquad \text{(3-2)}$$

With feasibility ratings from each participant (evaluator), aggregate TASC-feasibility ratings can be calculated by averaging participant ratings for each design. These scores were then used to rank the ideas according to their TASC-feasibility score by assigning a value of 1 (most feasible) to the design with the highest TASC-feasibility score, and 9 (least feasible) to the concept that had the lowest TASC-feasibility score. This was completed for the nine designs within each problem domain (milk frother, toothbrush, and texting).

TASC-creativity

Once the TASC-innovative and TASC-feasibility scores are calculated, the TASC-creativity metric can be computed. The TASC-creativity metric is meant to provide a global assessment of concept creativity because design creativity is often described as ideas that are both novel, and technically feasible [6, 9]. Specifically, the TASC-creativity rating is calculated by taking a direct sum of the TASC-innovative, and TASC-feasible ratings, i.e.,

$$C_{ijk} = I_{ijk} + F_{ijk}\ . \qquad \text{(3-3)}$$

With creativity ratings from each participant (evaluator), aggregate TASC-creativity ratings can be calculated by averaging participant ratings for each design. These scores were then used to rank the ideas according to their TASC-creativity score by assigning a value of 1 (most creative) to the design with the highest TASC-creativity score, and 9 (least creative) to the concept that had the lowest TASC-creativity score. This was completed for the nine designs within each problem domain (milk frother, toothbrush, and texting).

Perceived Novelty and Feasibility

Finally, in order to understand how design engineers perceive the novelty and feasibility of a candidate concept, each of the 27 design concepts was evaluated using 100-point evaluation scales in the second part of the study. This type of evaluation has been used in industry to help teams provide feedback and make decisions [89]. One hundred-point evaluation systems have also been utilized throughout the fields of psychology, education, and business to obtain feedback [90-92]. For this reason, it was utilized in this study as a subjective measure of design novelty and quality. This metric was purely the value each participant assigned for each concept's feasibility and novelty.

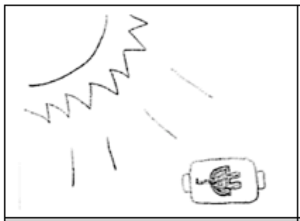
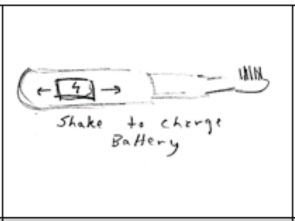
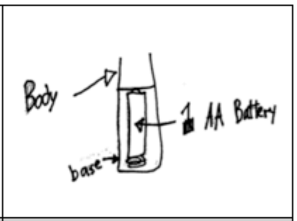
To provide an overall evaluation of perceived creativity, a perceived-creativity composite rating was also calculated by taking a summation of the novelty rating and the feasibility ratings provided by each participant for each concept that was evaluated. This composite rating was used to compare and contrast ratings and rankings of concepts. The perceived-creativity composite score P_{ijk} is calculated using

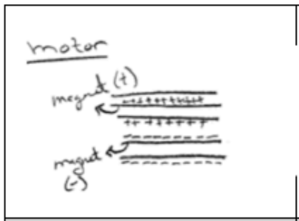
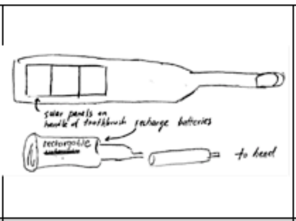
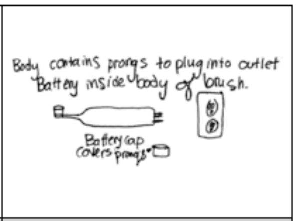
$$P_{ijk} = N_{ijk} + Q_{ijk}, \quad \text{(3-4)}$$

where N_{ijk} is the perceived-novelty rating for concept i from design problem j by participant k and where Q_{ijk} is the perceived-feasibility rating for concept i from design problem j by participant k.

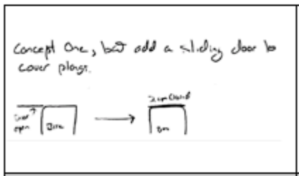
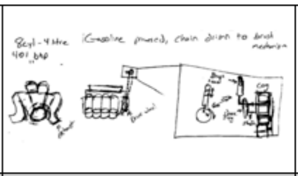
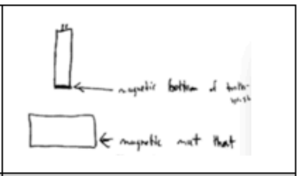
With creativity ratings from each participant (evaluator), aggregate perceived-creativity ratings can be calculated by averaging participant ratings for each design. These scores were then used to rank –order the ideas according to their perceived creativity score by assigning a value of 1 (most creative) to the design with the highest perceived creativity score and 9 (least creative) to the concept that had the lowest perceived creativity score. This was completed for the nine designs within each problem domain (milk frother, toothbrush and texting).

The nine "tooth brush" design sketches evaluated in this study are shown in Figure 4 to provide an example of the average scores obtained from each evaluation methods.

			
	Design 1	**Design 2**	**Design 3**
Perception	(M = 0.32, SD = 0.22)	(M = 0.66, SD = 0.23)	(M = 0.05, SD = 0.06)
TASC	(M = 0.34, SD = 0.14)	(M = 0.41, SD = 0.14)	(M = 0.28, SD = 0.06)
SVS	0	0.30	0.51

			
	Design 4	**Design 5**	**Design 6**
Perception	(M = 0.52, SD = 0.31)	(M = 0.47, SD = 0.22)	(M = 0.38, SD = .27)
TASC	(M = 0.49, SD = 0.25)	(M = 0.48, SD = 0.29)	(M = 0.34, SD = 0.17)
SVS	0.28	0.55	0.85

motor
magnet (+)
magnet (-)
solar panels on handle of toothbrush recharge batteries
rechargable
to head
Body contains prongs to plug into outlet
Battery inside body of brush.
Battery cap covers prongs

			
	Design 7	**Design 8**	**Design 9**
Perception	(M = 0.22, SD = .22)	(M = 0.57, SD = 0.30)	(M = 0.52, SD = 0.31)
TASC	(M = 0.25, SD = 0.05)	(M = 0.41, SD = 0.22)	(M = 0.43, SD = 0.25)
SVS	0.39	0.70	1

Concept One, but add a sliding door to cover plugs.
open
closed
Gasoline powered, chain drive to brush mechanism
magnetic bottom of toothbrush
magnetic mat that

Figure 4. Summary comparisons of design evaluations from "toothbrush" design problem.

5.0 Results and Discussion

Before analyzing the results with reference to our research questions, an inter-rater reliability analysis was completed to test the reliability of each method. Specifically, Cohen's Kappa was calculated for all metrics for both novelty and quality, see Figure 5. The results showed that all of the metrics achieved an inter-rater reliability of 0.7 or above, which is considered to be "substantial agreement" [93]. The following sections present the results of the remainder of our analysis in relation to our research hypotheses.

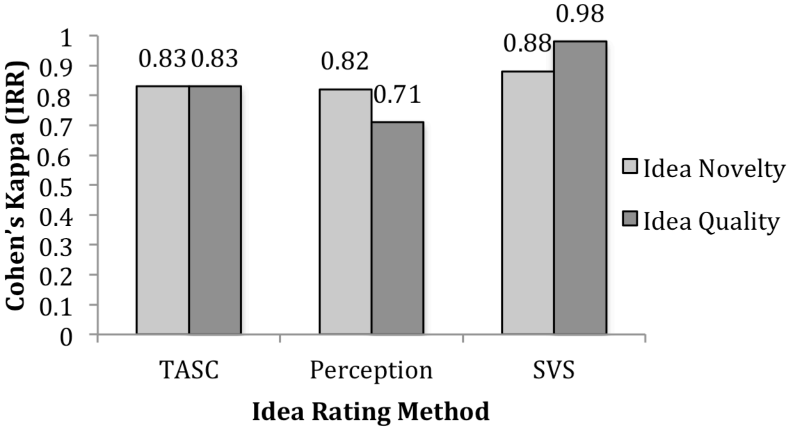


Figure 5. The inter-rater reliability (Kappa) for the idea rating methods used in the current study based on the 22 raters.

Do experts' and novices' perceptions of ideas differ in terms of design novelty, quality and overall creativity?

Our first research question sought to understand similarities and differences between expert and novice designers' perceptions of idea novelty, quality, and overall creativity. Specifically, our hypothesis was that expert and novice design engineers would evaluate design novelty in a similar light but diverge in their evaluations of design quality and thus their perception of a design's overall creativity. In order to answer this research question, a series of Spearman's Rank correlations were conducted between novices and experts perception of concept novelty, quality and creativity (6 total metrics). The two-tailed tests of significance indicated that a positive significant relationship between expert and novice perception of design concept novelty (r_s (27) = 0.741, $p < 0.01$), quality (r_s (27) = 0.749, $p < 0.01$) and creativity (r_s (27) = 0.861, $p < 0.01$).

Once it was identified that the expert and novice perceptions of these variables trended in the same direction through the correlation statistics, Cohen's Weighted Kappa was also calculated to determine the level of agreement on the rating given to each design concept (the Inter-rater reliability). The results revealed that moderate agreement on the expert and novices ratings of concept novelty (k= 0.51) and creativity (k = 0.65) but only slight agreement for concept quality ratings (k = 0.18) according to Landis's classification of kappa values [93].

These findings suggest that aggregate ratings from eleven untrained, novice designers can be used as a proxy for expert design ratings for overall design creativity and design novelty. This finding supports prior work in engineering design that found that aggregate scores of *40 highly-trained* novice raters can be used as a reliably proxy for an expert rater [94]. However, the novice designers in the current study received no training on the design tasks or rating scheme and our results indicated that only eleven raters are needed to mimic expert responses. This result contradicts prior research that has suggested that the limited experiences of novice designers have will also limit their case-based knowledge and thus their ability to effectively evaluate designs dissimilar from their experiences [95]. While the differences identified between the current study and prior research suggests an opportunity to dig deeper into the nuances of expert and novice evaluations, the results also suggest that an aggregate score of a few novice designers can be used to mimic expert responses.

Does the TASC method align with human perception or with the SVS method?

Our second research question sought to understand the similarities and differences between expert and novice designers' perception of creativity, our TASC method and evaluations using the SVS method. Specifically, our hypothesis was that our TASC method would tap into similar constructs of creativity used for perceived creativity and relative measures such as the SVS and thus would have some significant positive relationship with both measures. In this way our TASC method could be used to harness the benefits each of the prior methods and minimize possible experience biases.

In order to answer this research question, Cohen's weighted Kappa was conducted between the novice designers' ratings of idea creativity and the ratings from the TASC and SVS methods in order to determine the level of agreement on the rating given to each design concept (the inter-rater reliability), see Figure 6. The results revealed a moderate relationship between the novice perception and novice TASC ratings for the toothbrush (k = 0.88) and texting (k = 0.78) problems and a fair relationship for the milk frother problem (*k* = 0.35). This may be due to the participant's lack of familiarity with milk frothers; prior work has shown that case-based knowledge, although beneficial in most cases, can cause erroneous conclusions from raters when conditions are *not explicitly within the evaluator's perceived domain knowledge* [96, 97]. Novices are also likely to attribute judgments erroneously by linking design characteristics to prior experiences even if they are irrelevant to the design's feasibility [80]. It is also interesting to note that the results showed that the relationship between the novices' perception scores and the SVS ratings revealed only a fair agreement for the toothbrush (*k* = 0.33) and milk frother problems *(k*= 0.20) and only a slight agreement for the texting problem (*k* = 0.18). This finding indicates that the TASC method is tapping into more similar constructs of perceived design creativity than the SVS method.

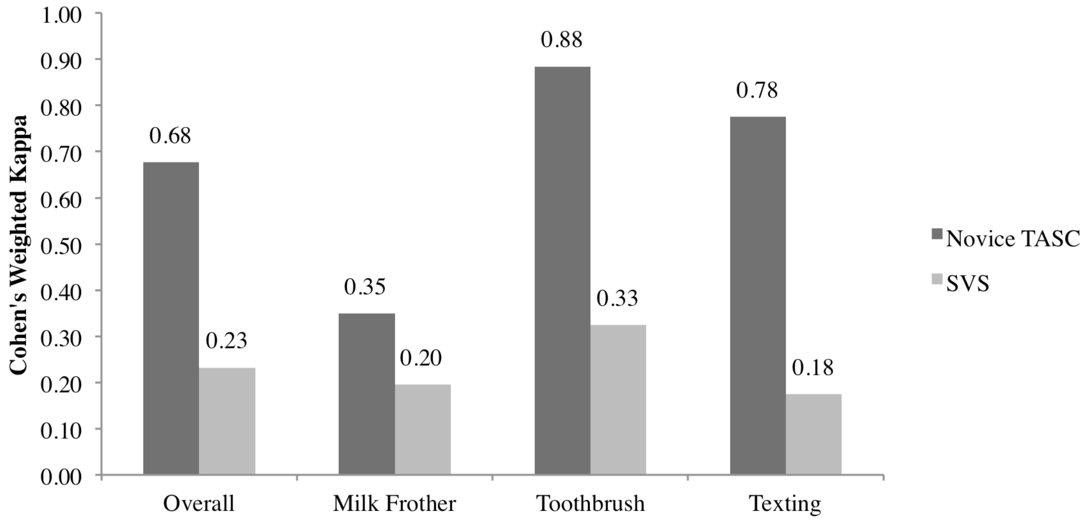


Figure 6: A summary of the Cohen's Weighted Kappa between novice rater's perception, novice TASC and SVS scores of all 27 designs.

In order to understand if a similar relationship exists with expert raters, Cohen's weighted kappa was also calculated between the expert designers' perception of creativity, expert TASC scores, and the SVS method, see Figure 7. The results revealed a moderate relationship between the experts' perception and expert TASC ratings for all of the design problems; milk frother (*k* = 0.58), toothbrush (*k* = 0.48) and texting (*k* = 0.48). However, the relationship between the expert perception scores and the SVS ratings revealed only moderate agreement for the milk frother problem (*k* = 0.43) and only fair agreement for the toothbrush *(k*= 0.33) and texting (*k* = 0.40) problems. Like the results from the novice designers, this result also indicates that the TASC method is tapping into more similar constructs of perceived design creativity than the SVS method.

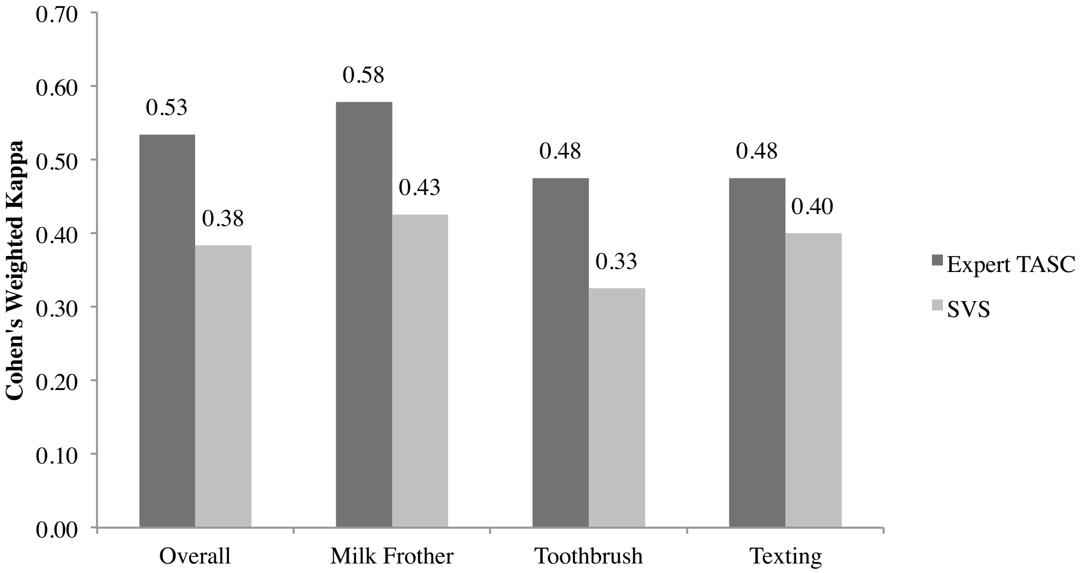


Figure 7: A summary of the Cohen's Weighted Kappa analysis between expert rater's perception, expert TASC and SVS scores of all 27 designs. ** Significant at *p* < 0.01, * *p* < 0.05

Finally, in order to understand how the TASC method compares to the 'gold standard' in the field (the SVS method), Cohen's weighted kappa was computed between the SVS ratings and both expert and novice TASC ratings. The results revealed only a 'fair' agreement between the SVS method and expert (*k* = 0.28) and novice (*k* = 0.28) TASC ratings. These results indicate that while there are some similarities between these measures, they produce different ratings of concept creativity.

The results from these tests support our hypothesis that our TASC method is tapping into similar constructs of creativity as human perception for both expert and novice designers. In addition, the results indicate that the TASC rating can be used as a better approximation of perceived ratings of design creativity than the SVS method for both novice and expert evaluators. This finding also suggests the SVS rating methods may not be tapping into a similar view of creativity as perceived by experts in product design or as measured by the TASC tool. This could be attributed to the fact that the SVS method is based on the relative creativity of the ideas in the set being evaluated while the TASC and perception measures are based on historical creativity, or the fundamentally novelty of an idea with respect to the whole of human history [47] which limits the utility of SVS. In addition, the SVS method has been criticized for the extensive rater training needed to combat low inter-rater reliabilities and difficulties interpreting multiple metrics simultaneously [39, 40]. Therefore, the finding from the current studying is promising for the development of more absolute and *global* measures of design creativity with minimal training intervention. In this way, the TASC supports the effort to allow the same metric and framework to ultimately evaluate different design problems on a global scale.

Does the TASC and SVS method align with expert human perception and can TASC be used as a proxy for expert ratings?

Our final research question was developed to identify if, or how well, the TASC metric and novice perception can be used as proxies for expert ratings. Our hypothesis was that our TASC method, when used by novices, would produce evaluations comparable to those by experts due to the use of adjective selections that could reduce experience dependence. In order to answer this question, a series of Cohen's Weighted Kappa calculations were conducted. Summaries of this analysis, shown in Figure 8, are provided below.

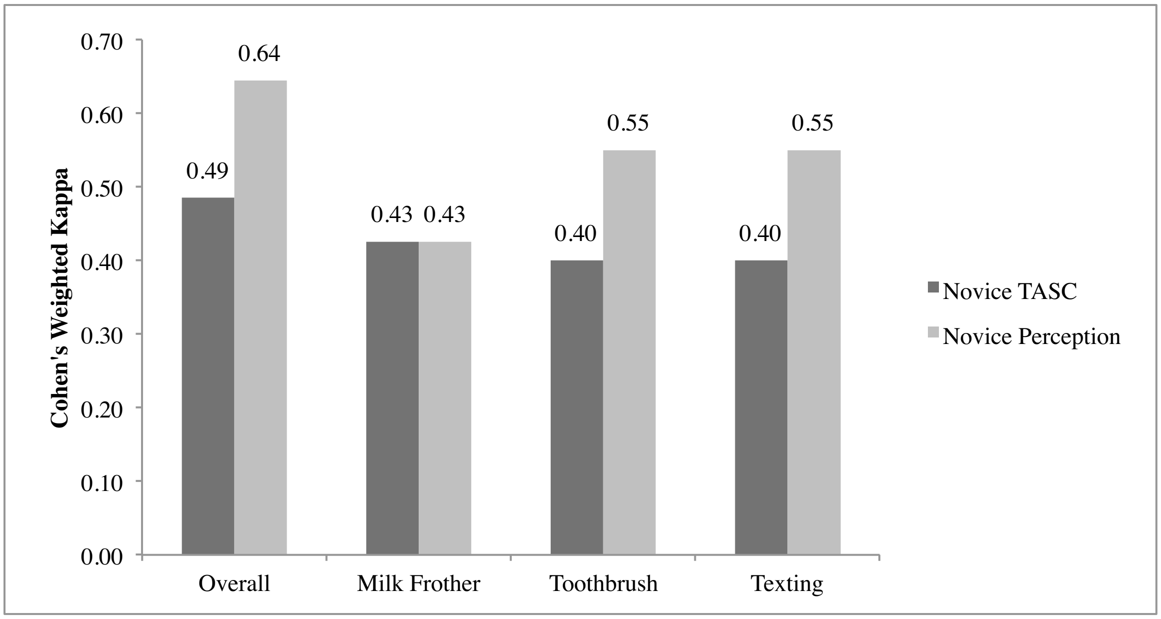


Figure 8: A summary of the Cohen's Weighted Kappa calculations between expert perception and both novice TASC and novice perception scores by design problem.

Our results showed a moderate relationship between the expert and novice's perception for all of the design problems: milk frother (*k* = 0.43), toothbrush (*k* = 0.55) and texting (*k* = 0.55). This result demonstrates that the openness of the design problem had no significant impact on the strength of the positive correlation between expert and novice ratings of design creativity. The novice's TASC scores also had a moderate relationship with all three of the design problems: milk frother (*k* = 0.43), toothbrush (*k* = 0.40) and texting (*k* = 0.40). However, the kappa coefficients were not as high for the novice TASC scores as they were for the novice perception scores.

These findings demonstrate that while the TASC method shows promise to be used as proxy for expert ratings of design concept creativity, the average ratings of eleven novice designers' perception of creativity is actually more effective of a measurement. Interestingly, this argument holds true regardless of the “openness” of the design concept being evaluated. These findings neither support nor reject our hypothesis that our TASC method can be used for a proxy for expert perception. However, they do show that there is potential for our TASC method to reduce experience biases and enable novices to obtain expert-level evaluations.

By using words selected from a predefined set, the TASC method provides a streamlined framework for diagnostic feedback to designers as words selections have in Kansei engineering [65] and the desirability toolkit [26]. Although our TASC method had a moderate relationship with all three design problems, future work will be required to further improve the accuracy of this method. This could be possible through the development of a crowd-sourced semantic similarity index using Amazon Mechanical Turk [68] that could provide word weightings that are more in line with human intuition. This could further the push for creativity assessment tools that bridge the gap between fast human perception and the repeatability of the SVS method. The results from comparing novice perception to expert perception showed strong relationships between the two groups. These findings support the effectiveness of novice evaluators beyond prior use in crowdsourcing research [94]. It also strengthens the argument for utilizing novice evaluators in design evaluation tools as low cost and more accessible alternative to expert evaluators.

3.6 Impetus for Engineering Design Education and Research

The main goals of this research were to further our understanding of how expert and novice perceptions of creativity relate to other measures and investigate the development and use of our TASC in comparison to human perception and prior creativity metrics. Our results revealed the following key results:

1. Expert and novice raters were in strong agreement with their perceptions of creativity in concept designs regardless of the openness of the design problem;
2. Our TASC method was able to tap into similar constructs of expert and novice perceptions of creativity in concept design;
3. Aggregate scores of 11 untrained novice designers can be used as a proxy for expert ratings irrespective of design problem openness; and
4. While there is potential for using our TASC method with novice raters to achieve expert level feedback, more work is needed to refine the method to improve its utility over human perception.

These results have several important implications for engineering design and computational design creativity systems in education and industry. First, the results show that despite their varying levels of experience, experts and novices are able to reach similar conclusions about a design's creativity rating. Our results align with prior research in design expertise and crowd-sourced design that suggests that novices with minimal training can be used as a proxy for expert feedback [55, 98]. However, there was no training involved in our study which greatly improves its utility as an efficient evaluation method.

While it may be powerful to have numerous evaluators in product design due to the law of large numbers, our results have shown that even with 11 expert and 11 novice evaluators, we were able to obtain significant ratings. So, despite prior works in support of crowd sourcing especially for novices in product design [99, 100], numerous evaluators may not be necessary to effectively evaluate design creativity. This means that time and resources can be better allocated towards design efforts. This finding also enables the use of creativity evaluation methods such as our TASC method to streamline the evaluation process for industry and within education. Building on education, it might be possible for students to evaluate the designs within a classroom setting without finding overly confined evaluator groups or spending money.

In addition to highlighting the potential of novice evaluations of concept creativity, the results of this study establish the reality of computational design-creativity systems as a means to substantiate creative designs in the selection process. Prior studies in engineering design and psychology have shown that few creative designs actually survive the concept selection process due to biases that stigmatize creativity [101, 102]. Our TASC method provides a framework in which qualitative data becomes multifaceted during the design process. At first glance, the words can be analyzed on their own for how the designer’s message has been communicated through the sketch. The assignment of semantic weights provides quantitative values that can be used to draw comparisons between the designs and substantiate design decision. Thus, the design evaluation method developed in this research pushes for quality as well as creativity within the design process.

3.7 Limitations and Future Work

While the current study highlighted the development of computational creativity evaluation tools in concept selection and identified the use of such tools with novice raters, there are several important limitations that should be noted. The most important limitation is that this study was developed using words that originated from the Desirability Toolkit developed by Benedek and Miner [26] to obtain user feedback on desirability. Although many of the words used within the Desirability Toolkit are applicable within engineering design, there is an opportunity for future work to tune the word list more appropriately. For example, words can be borrowed from affective and Kansei engineering and implemented in the TASC framework with relative ease. There is also an opportunity to adjust the word selections to better suit other areas of design. This would help develop adjectives for use in our TASC methodology that have more distinct innovation and feasibility ratings. This is important because the polarity of the innovation and feasibility weights in our current word set our similar.

In addition to the word selection list, there is an opportunity to explore more advanced measures of word relatedness within the TASC method. The proliferation of natural language processing techniques and machine learning technologies has the potential to increase the correlation between computed word relatedness and human perceived word relatedness. The Java class DISCO [87] was used to compute semantic similarity in the current study due to its accessibility and strong relationship with human perception among other freely available solutions. We are also interested in developing customized word relatedness indexes based on human feedback using crowdsourcing tools such as Mechanical Turks as supported in prior studies [55, 103]. However, further experimental investigations on this topic can be implemented within our TASC methodology with relative ease.

Finally, the current study identified that the TASC method can be useful as a proxy for expert level feedback, novice perception aligns more strongly with expert opinion. This finding identifies that while there is a potential for using the TASC method with novice raters to achieve expert level feedback, further experimentation is needed to understand the factors that impact the utility of this approach. In addition, our classification of an expert was based on a two-pronged approach: more than three years of design experience and a self-classification of an expert on a five point Likert scale. While this was used to insure they were an expert in engineering design, they did not receive any training to familiarize themselves with the problem domains being explored meaning they were not necessarily experts in those particular problem topics. These limitations call for future work that is geared at understanding the impact of the design task, the designers experience level, familiarity with the task domain and the design domain on the utility of the TASC approach.

4.0 Conclusions

The main goal of this study was to investigate the utility of our TASC method and explore the relationship between evaluator experience and various design concept creativity evaluation methods. To meet this goal, quantitative and qualitative data were collected and analyzed from a controlled study utilizing an online questionnaire with expert and novice design engineers. Overall, the results of this study show that novice and expert evaluators perceive concept creativity in a similar light and demonstrate that it may be possible to utilize this similarity to reduce the costs and limitations of using expert evaluations in the concept evaluation process. Our results also showed support for using computational design-creativity tools such as TASC to assess creativity without training participants. These types of tools have an opportunity to simplify the concept evaluation process and make it accessible and practical to assess concepts in industry and academia. Our results are used to provide directions for future research and provide recommendations for design evaluation that support creativity throughout the design process.

5.0 Acknowledgments

This material is based upon work supported by The National Science Foundation under Grant No. 1351493. We would also like to thank our undergraduate research assistants and our participants for their aid in this research project

6.0 References

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

[2] Osborn, A. F., 1963, Applied Imagination: Principles and procedures of creative thinking, Scribeners and Sons, New York.
[3] Yang, M. C., 2009, "Observations on concept generation and sketching in engineering design," Research in Engineering Design, 20(1), pp. 1-11.
[4] Chulvi, V., Gonzalez-Cruz, M. C., Mulet, E., and Aguilar-Zambrano, J., 2012, "Influence of type of idea-generation method on the creativity of solutions," Research in Engineering Design, 24(11), pp. 33-41.
[5] Yilmaz, S., Seifert, C. M., and Gonzalez, R., 2012, "Design Heuristics: Cognitive Strategies for Creativity in Idea Generation," Design Computing and Cognition, J. S. Gero, ed., pp. 35-53.
[6] Shah, J. J., Vargas-Hernandez, N., and Smith, S. M., 2003, "Metrics for Measuring Ideation Effectiveness," Design Studies, 24(1), pp. 111-134.
[7] Rietzchel, E. F., Nijstad, B. A., and Stroebe, W., 2006, "Productivity is not enough: a comparison of interactive and nominal groups in idea generation and selection," Journal of Experimental Social Psychology, 42(2), pp. 244-251.
[8] Snider, C., Cash, P., Dekoninck, E., and Culley, S., "Variation in creative behaviour during the later stages of the design process," Proc. ICDC2012: The 2nd International Conference on Design Creativity, University of Bath, pp. 147-156.
[9] Kudrowitz, B. M., and Wallace, D., 2013, "Assessing the quality of ideas from prolific, early-stage product ideation," Journal of Engineering Design, 24(2), pp. 120-139.
[10] Pahl, G., and Beitz, W., 1984, Engineering Design, The Design Council, London.
[11] Fishburn, P. C., 1970, "Utility theory for decision making," DTIC Document, Mclean, VA.
[12] Von Neumann, J., and Morgenstern, O., 1953, Theory of Games and Economic Behavior: 3d Ed, Princeton University Press.
[13] Marsh, E. R., Slocum, A. H., and Otto, K. N., 1993, "Hierarchical decision making in machine design," MIT Precision Engineering Research Center.
[14] Saaty, T. L., 1988, What is the analytic hierarchy process?, Springer Berlin Heidelberg.
[15] Saaty, T. L., 1980, The Analytic Hierarchy Process, McGraw Hill, New York, NY.
[16] Pugh, S., 1991, Total Design, Addison-Wesley.
[17] Pugh, S., 1991, Total design: integrated methods for successful product engineering, Addison-Wesley, Workingham.
[18] Pugh, S., 1981, "Concept selection- a method that works," International Conference of Engineering DesignRome, Italy, pp. 497-506.
[19] Ter Harr, S., Clausling, D., and Eppinger, S., 1993, "Integration of Quality Function Deployment in the Design Structure Matrix," Cambridge, MA.
[20] Hauser, J., and Clausing, D., 1996, "The house of quality," IEEE Engineering Management Review, 24(1), pp. 24-32.
[21] Ross, T., 1995, Fuzzy Logic with Engineering Applications, McGraw Hill.
[22] Thurston, D. L., and Carnahan, J. V., 1992, "Fuzzy Ratings and Utility Analysis in Preliminary Design Evaluation of Multiple Attributes," Journal of Mechanical Design, 114, pp. 648-658.

[23] Okugan, G. E., and Tauhid, S., 2008, "Concept Selection Methods – A Literature Review from 1980 to 2008," International Journal of Design Engineerin, 1(3), pp. 243-277.
[24] Genco, N., Holtta-Otto, K., and Seepersad, C. C., 2012, "An Experimental Investigation of the Innovation Capabilities of Undergraduate Engineering Students," Journal of Engineering Education, 101(1), pp. 60-81.
[25] Gray, D. E., Brown, S., and James, M., 2010, "NUF Test," Gamestorming:, O'Reilly Media, Sebastopol, CA.
[26] Benedek, J., and Miner, T., 2002, "Measuring Desirability: New methods for evaluating desirability in a usability lab setting," Proceedings of Usability Professionals Association, 2003, pp. 8-12.
[27] 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.
[28] 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.
[29] Amabile, T., 1982, "Social psychology of creativity: A consensusual assessment technique," Journal of Personality and Social Psychology, 43, pp. 997-1013.
[30] Brown, D., 2013, "Developing computational design creativity systems," International Journal of Design Creativity and Innovation, 1(1), pp. 43-55.
[31] 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.
[32] 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.
[33] Sternberg, R. J., and Lubart, T. I., 1999, "The concept of creativity: Prospects and paradigms," Handbook of creativity, 1, pp. 3-15.
[34] Fischer, G., 2013, "Learning, Social Creativity, and Cultures of Participation," Learning and Collective Creativity: Activity-Theoretical and Sociocultural Studies, p. 198.
[35] Martin, M. W., 2006, "Moral creativity in science and engineering," Science and engineering ethics, 12(3), pp. 421-433.
[36] Nelson BA, Y. J., 2009, "Refined metrics for measuring ideation effectiveness," Design Studies, 30, pp. 737-743.
[37] Chulvi, V., Mulet, E., Chakrabarti, A., López-Mesa, B., and González-Cruz, C., 2012, "Comparison of the degree of creativity in the design outcomes using different design methods," Journal of Engineering Design, 23(4), pp. 241-269.
[38] Lopez-Mesa, B., Mulet, E., Vidal, R., and Thompson, G., 2011, "Effects of additional stimuli on idea-finding in design teams," Journal of Engineering Design, 22(1), pp. 31-54.
[39] Nelson, A., Matz, M., Chen, F., Siddharthan, K., Lloyd, J., and Fragala, G., 2006, "Development and evaluation of a multifaceted ergonomics program to prevent injuries

associated with patient handling tasks," International journal of nursing studies, 43(6), pp. 717-733.
[40] Srivathsavai, R., Genco, N., Holtta-Otto, K., and Seepersad, C., 2010, "Study of Existing Metrics Used in Measurement of Ideation Effectiveness," ASME 2010 International Design Engineering Technical Conferences & Computers and Information Engineering ConferenceMontreal, Quebec, Canadac.
[41] Amabile, T. M., 1983, "The social psychology of creativity: A componential conceptualization," Journal of personality and social psychology, 45(2), pp. 357-376.
[42] 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, pp. 120-151.
[43] Bedell-Avers, K., Hunter, S. T., Angie, A. D., Eubanks, D. L., and Mumford, M. D., 2009, "Charismatic, ideological, and pragmatic leaders: An examination of leader–leader interactions," The Leadership Quarterly, 20, pp. 299-315.
[44] Hunter, S. T., Bedell, K. E., Ligon, G. S., Hunsicker, C. M., and Mumford, M. D., 2008, "Applying multiple knowledge structures in creative thought: Effects on idea generation and problem-solving," Creativity Research Journal, 20, pp. 137-154.
[45] Lopez-Mesa, B., and Bylund, N., 2011, "A study of the use of concept selection methods from inside a company," Research in Engineering Design, 22(1), pp. 7-27.
[46] Brown, D. C., "Problems with the Calculation of Novelty Metrics," Proc. Proc. Design Creativity Workshop, 6th Int. Conf. on Design Computing and Cognition (DCC'14).
[47] Fischer, G., 2013, "Learning, Social Creativity, and Cultures of Participation," Learning and Collective Creativity: Activity-Theoretical and Sociocultural Studies, A. Sannino, and V. Ellis, eds., Taylor & Francis/ Routledge, New York, NY.
[48] Maher, M. L., and Fisher, D. H., "Using AI to evaluate creative designs," Proc. 2nd International Conference on Design Creativity, Glasgow, UK.
[49] Gero, J. S., 1990, "Design Prototypes: A knowledge representation schema for design," AI Magazine, pp. 22-36.
[50] Gero, J. S., and Kannengiesser, U., 2007, "Locating creativity in a framework of designing for innovation," Trends in computer aided innovation, Springer, pp. 57-66.
[51] Shiffrin, R. M., and Schneider, W., 1977, "Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory," Psychological review, 84(2), p. 127.
[52] Baddeley, A. D., 2002, "Is working memory still working?," European psychologist, 7(2), p. 85.
[53] Cardoso, C., Goncalves, M., and Badke-Schaub, P., 2012, "Searching for Inspiration During Idea Generation: Pictures or Words?," International Design ConferenceCroatia, pp. 1831-1840.
[54] Licuanan, B. F., Dailey, L. R., and Mumford, M. D., 2007, "Idea evaluation: Error in evaluating highly original ideas," The Journal of Creative Behavior, 41(1), pp. 1-27.
[55] Green, M., Seepersad, C., and Hölttä-Otto, K., 2014, "Crowd-sourcing the Evaluation of Creativity in Conceptual Design: A Pilot Study," ASME IDETC Design Theory and Methodology Conference, DETC2014-34434, Buffalo, NY.
[56] Cross, N., 2004, "Expertise in design:an overview," Design Studies, 25(5), pp. 427-441.

[57] Atman, C. J., Adams, R. S., Cardella, M. E., Turns, J., Mosborg, S., and Saleem, J., 2007, "Engineering design processes: A comparison of students and expert practitioners," Journal of Engineering Education, 96(4), pp. 359-379.
[58] Worsley, M., and Blikstein, P., "What's an Expert? Using Learning Analytics to Identify Emergent Markers of Expertise through Automated Speech, Sentiment and Sketch Analysis," Proc. EDM, pp. 235-240.
[59] Ericsson, K. A., and Lehmann, A. C., 1996, "Expert and exceptional performance: Evidence of maximal adaptation to task constraints," Annual review of psychology, 47(1), pp. 273-305.
[60] Jiao, J. R., Zhang, Y., and Helander, M., 2006, "A Kansei mining system for affective design," Expert Systems with Applications, 30(4), pp. 658-673.
[61] Han, S. H., and Hong, S. W., 2003, "A systematic approach for coupling user satisfaction with product design," Ergonomics, 46(13-14), pp. 1441-1461.
[62] Chuang, M.-C., and Ma, Y.-C., 2001, "Expressing the expected product images in product design of micro-electronic products," International Journal of Industrial Ergonomics, 27(4), pp. 233-245.
[63] Korpershoek, H., Kuyper, H., Werf, G. v. d., and Bosker, R., 2010, "Who ‘fits’ the science and technology profile? Personality differences in secondary education," Journal of Research in Personality, 44(5), pp. 649-654.
[64] Choi, K., and Jun, C., 2007, "A systematic approach to the Kansei factors of tactile sense regarding the surface roughness," Applied Ergonomics, 38(1), pp. 53-63.
[65] Childs, T., Agouridas, V., Barnes, C., and Henson, B., 2006, "Controlled appeal product design: a life cycle role for affective (Kansei) engineering," Proceedings of LCE2006, pp. 537-542.
[66] Shah, J. J., Millsap, R. E., Woodward, J., and Smith, S. M., 2012, "Applied Tests of Design Skills, Part 1: Divergent Thinking," Journal of Mechanical Design, 134(2), pp. 021005-021005.
[67] Rogers, E., 1995, "M.(1995). Diffusion of innovations," The Free Press, New York.
[68] Bollegala, D., Matsuo, Y., and Ishizuka, M., 2011, "A web search engine-based approach to measure semantic similarity between words," Knowledge and Data Engineering, IEEE Transactions on, 23(7), pp. 977-990.
[69] Jin, X., and Mobasher, B., "Using semantic similarity to enhance item-based collaborative filtering," Proc. Proceedings of The 2nd IASTED International Conference on Information and Knowledge Sharing, pp. 1-6.
[70] Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S., "A word at a time: computing word relatedness using temporal semantic analysis," Proc. Proceedings of the 20th international conference on World wide web, ACM, pp. 337-346.
[71] Prabhakaran, R., Green, A. E., and Gray, J. R., 2013, "Thin slices of creativity: Using single-word utterances to assess creative cognition," Behavior research methods, pp. 1-19.
[72] Fabiani, M., and Donchin, E., 1995, "Encoding processes and memory organization: a model of the von Restorff effect," Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(1), p. 224.
[73] Snyder, K. A., Blank, M. P., and Marsolek, C. J., 2008, "What form of memory underlies novelty preferences?," Psychonomic bulletin & review, 15(2), pp. 315-321.

[74] Anderson, C. J., Glassman, M., McAfee, R. B., and Pinelli, T., 2001, "An investigation of factors affecting how engineers and scientists seek information," Journal of Engineering and Technology Management, 18(2), pp. 131-155.
[75] Peracchio, L. A., and Tybout, A. M., 1996, "The moderating role of prior knowledge in schema-based product evaluation," Journal of Consumer Research, pp. 177-192.
[76] Kurdrowitz, B., and Dippo, C., 2013, "Getting to the novel ideas: exploring the altenative uses test of divergent thinking," ASME Design Engineering Technical ConferencesPortland, OR.
[77] Nikander, J. B., Liikkanen, L. A., and Laakso, M., 2014, "The preference effect in design concept evaluation," Design Studies, 35(5), pp. 473-499.
[78] Lu, C.-C., and Luh, D.-B., 2012, "A Comparison of Assessment Methods and Raters in Product Creativity," Creativity Research Journal, 24(4), pp. 331-337.
[79] Von Hippel, E., 1986, "Lead users: a source of novel product concepts," Management science, 32(7), pp. 791-805.
[80] Dailey, L., and Mumford, M. D., 2006, "Evaluative aspects of creative thought: Errors in appraising the implications of new ideas," Creativity Research Journal, 18(3), pp. 385-390.
[81] Toh, C., and Miller, S. R., 2013, "Product Dissection or Visual Inspection? The Impact of Designer-Product Interactions on Engineering Design Creativity," ASME Design Engineering Technical ConferencesPortland, OR.
[82] Toh, C. A., Miller, S. R., and Kremer, G. E., 2012, "Mitigating Design Fixation Effects in Engineering Design Through Product Dissection Activities," Design Computing and CognitionCollege Station, TX, p. n. pag.
[83] Miller, S. R., Bailey, B. P., and Kirlik, A., 2014, "Exploring the Utility of Bayesian Truth Serum for Assessing Design Knowledge," Human-Computer Interaction, 29(5-6), pp. 487-515.
[84] Toh, C., Miller, S., and Okudan Kremer, G., 2012, "Mitigating Design Fixation Effects in Engineering Design Through Product Dissection Activities," Design Computing and CognitionCollege Station, TX.
[85] Williams, D., Kelly, G., and Anderson, L., "MSN 9: new user-centered desirability methods produce compelling visual design," Proc. CHI'04 Extended Abstracts on Human Factors in Computing Systems, ACM, pp. 959-974.
[86] Kolb, P., "Experiments on the difference between semantic similarity and relatedness," Proc. Proceedings of the 17th Nordic Conference on Computational Linguistics-NODALIDA’09.
[87] Kolb, P., 2008, "Disco: A multilingual database of distributionally similar words," Proceedings of KONVENS-2008, Berlin.
[88] 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.
[89] Wells, J. D., Campbell, D. E., Valacich, J. S., and Featherman, M., 2010, "The effect of perceived novelty on the adoption of information technology innovations: a risk/reward perspective," Decision Sciences, 41(4), pp. 813-843.

[90] Endicott, J., Spitzer, R. L., Fleiss, J. L., and Cohen, J., 1976, "The Global Assessment Scale: a procedure for measuring overall severity of psychiatric disturbance," Archives of General Psychiatry, 33(6), p. 766.
[91] Chiou, C.-F., Hay, J. W., Wallace, J. F., Bloom, B. S., Neumann, P. J., Sullivan, S. D., Yu, H.-T., Keeler, E. B., Henning, J. M., and Ofman, J. J., 2003, "Development and validation of a grading system for the quality of cost-effectiveness studies," Medical Care, 41(1), pp. 32-44.
[92] Joyce, M., and Kirakowski, J., 2014, "Measuring Confidence in Internet Use: The Development of an Internet Self-efficacy Scale," Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience, Springer, pp. 250-260.
[93] Landis, J. R., and Koch, G. G., 1977, "The measurement of observer agreement for categorical data," biometrics, pp. 159-174.
[94] Green, M., Seepersad, C. C., and Hölttä-Otto, K., 2014, "Crowd-sourcing the evaluation of creativity in conceptual design: a pilot study," Proceedings of the ASME 2014 International Design Engineering Technical ConferencesBuffalo, NY.
[95] Weekley, J. A., and Gier, J. A., 1989, "Ceilings in the reliability and validity of performance ratings: The case of expert raters," Academy of Management Journal, 32(1), pp. 213-222.
[96] Chan, S., 1982, "Expert judgments made under uncertainty: Some evidence and suggestions.," Social Science Quarterly, 63, pp. 428-444.
[97] Anderson, J. R., 1987, "Skill acquisition: Compilation of weak-method problem situations," Psychological review, 94(2), p. 192.
[98] Poetz, M. K., and Schreier, M., 2012, "The value of crowdsourcing: can users really compete with professionals in generating new product ideas?," Journal of Product Innovation Management, 29(2), pp. 245-256.
[99] Wu, W., Luther, K., Pavel, A., Hartmann, B., Dow, S., and Agrawala, M., 2013, "CrowdCritter: Strategies for Crowdsourcing Visual Design Critique."
[100] Burnap, A., Ren, Y., Papalambros, P. Y., Gonzalez, R., and Gerth, R., 2013, "A simulation based estimation of crowd ability and its influence on crowdscourced evaluation of design concepts," ASME Design Engineering Technical ConferencesPortland, OR.
[101] Rietzschel, E., BA Nijstad, and W. Stroebe, 2010, "The selection of creative ideas after individual idea generation: choosing between creativity and impact.," British Journal of Psychology, 101(1), pp. 47-68.
[102] Mueller, J. S., Melwani, S., and Goncalo, J. A., 2011, "The Bias Against Creativity: Why People Desire But Reject Creative Ideas," Psychological Science, 2011, p. 0956797611421018.
[103] Kudrowitz, B. M., and Wallace, D., 2012, "Assessing the quality of ideas from prolific, early-stage product ideation," Journal of Engineering Design, 24(2), pp. 120-139.

Appendix: Design tasks used in prior studies for idea generation activities.

Portable Electric Toothbrush Power Design Task

Upper management has put your team in charge of developing a concept for a *new innovative power mechanism for a portable electric toothbrush.* Electric toothbrushes are popular personal devices used for dental hygiene. The advantages of an electrical toothbrush are many, including improved cleaning, ease of use, and other additional features. However, since electric toothbrushes require energy to function, this limits the portability of this device compared to manual toothbrushes.

Once again, the goal is to *develop concepts for a new,* ***innovative*** *power mechanism for a portable electric toothbrush. This product should be able to be used by the consumer with minimal instruction.*

Milk Frother Design Task

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. As such, devices that froth milk can also be used in a number of other applications, such as for whipping cream, blending drinks, emulsifying salad dressing, and many others. This design your team develops 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.

Once again, the goal is to *develop concepts for a new,* ***innovative*** *product that can froth milk in a short amount of time. This product should be able to be used by the consumer with minimal instruction.*

Mobile Device Solutions for Reducing Pedestrian Accident Rates Design Task

Upper management has put your team in charge of developing a concept for a *new innovative product or technology that reduces student accident rates associated with using a cell phone mp3 player while walking around campus.* There has been an increase in student accidents on campus in recent years from student's texting and/ or talking on cellphones or listening to music using earphones while walking around campus. While using these devices, students become distracted, and can trip, fall or even collide into obstacles. In fact, in 2008, over 1,000 pedestrians visited emergency rooms due to accidents from using these devices while walking. There are reports of concussions, sprained ankles, broken appendages and even fatalities from these accidents. These numbers do not include the countless number of unreported incidents involving walking into something (i.e. a parked car) without an ER visit. This increase in accidents has been substantial on college campuses because of the number of students on campus and the increased usage of mobile devices (listening to music, texting ,and talking) all of which are distracting.

Once again, the goal is to *develop concepts for a new innovative product or technology that reduces student accident rates associated with walking and using a cell phone or mp3 player while walking around campus.*