The Pennsylvania State University

The Graduate School

College of Engineering

TOWARD THE DEVELOPMENT OF METHODS TO SUPPORT CREATIVE CONCEPT SELECTION IN ENGINEERING DESIGN

A Thesis in

Industrial Engineering

by

Christopher A. Gosnell

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2014

The thesis of Christopher A. Gosnell was reviewed and approved* by the following:

Scarlett. R. Miller
Assistant Professor of Mechanical and Industrial Engineering
Thesis Adviser

Sven G. Bilén
Associate Professor of Engineering Design, Electrical Engineering, and Aerospace Engineering

Paul M. Griffin
Professor of Industrial Engineering
Head of the Industrial and Manufacturing Department

*Signatures are on file in the Graduate School

ABSTRACT

Engineering design idea-generation sessions often result in dozens, if not hundreds, of ideas. These ideas must be quickly evaluated and filtered in order to select a few candidate concepts to move forward in the design process. While creativity is often stressed in the conceptual phases of design, it receives little attention in these later phases – particularly during concept selection. This is largely because there are no methods for quickly rating or identifying worthwhile creative concepts during this process. Therefore, the purpose of this thesis was to develop and test a novel method for evaluating the creativity and feasibility of design concepts. The first phase of this research involved the creation of a creativity evaluation methodology that utilizes word selections and semantic similarity to quickly and effectively evaluate candidate concepts. To test its utility during concept selection, an empirical study with ten engineering designers was completed. The results revealed that our methodology and rating system could be used as a proxy for measuring design creativity regardless of the openness of the design task. The second phase of this thesis sought to investigate the impact of decision-making bias during concept selection and to refine our design creativity evaluation methodology. In order to accomplish this, an online questionnaire was developed and administered to 11 expert and 11 novice design engineers. The results from this study supported the use of our creativity evaluation methodology with both expert and novice raters for early design creativity. It also contributes to our understanding of experience bias in creativity evaluation and provided a framework for computational design creativity systems. The final phase of this thesis sought to explore the use of the creativity assessment methodologies within a classroom setting and to compare the results from team evaluations and individual team member evaluations. During this phase, our creativity evaluation methodology was developed into a printable toolkit to help students evaluate designs for an in-class project. In this study, 32 students in teams of four utilized the toolkit to analyze design sketches from their team members. The results from this study showed significant relationships between individual and team perception based evaluations that utilized sliding scales and group discussion and our adjective selection method. However, the widely adopted creativity evaluation method from academia that compared designs feature by feature appeared to be measuring something else entirely. The results from this thesis are used to develop recommendations for the use of creative concept selection tools in engineering design. In addition, the results are extrapolated to create recommendations for an online web application geared enhancing the usability and accessibility of concept selection tools for both academic and industrial implementation.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor, Dr. Scarlett Miller for identifying my potential early on and for constantly pushing me to be confident in my abilities and to seek perfection. Her dedication towards developing the minds and lives of students is without comparison. In the last year and a half, I have learned more from her than most could have hoped for in their entire academic lives. She has taught me how to seize every opportunity available and to maximize the potential of every day. She has been more than just an advisor. She has been my mentor, teacher and overall life coach. In this way, she has shown me how push others and myself beyond our self-perceived limits while maintaining balanced lives. I am forever grateful for her time and patience while working with me and I am truly honored to have worked with such an inspiring individual.

Finally, I would like to thank all members from the Brite Lab for their constant support in both my research and life in State College. They have been my sounding board for numerous practice presentations and have been key to the success of my research studies. I would like to especially thank my fellow graduate students Christine Toh and Elizabeth Starkey for being available at a moment's notice to provide guidance and feedback both in my academic and professional development. I am also thankful to Dr. Sven Bilén for accepting me into the Integrated Design Solutions (IDS) group on campus and helping me to gain interdisciplinary experience.

Chapter 1

Introduction

Design generation tools and methodologies are plentiful and have given design teams the ability to generate tens, if not hundreds, of ideas in a relatively short period of time [1]. While prior studies have shown that there is a positive relationship between the number of designs generated and the likelihood of creative ideas being included in the set [2], this influx of design ideas can be problematic for design teams that need to effectively select the most creative designs to develop. In an increasingly competitive marketplace, design creativity can mean the difference between market leading products such as the Toyota Prius^{TM} and the lesser known Honda Civic^{TM} hybrid [3, 4].

Although a number of decision-making tools are available to help designers with their selections [5-8], these tools are often complicated and require design parameters that may not yet exist in the early phases of design. In addition, these methods, as well as the designers themselves, have a tendency to overlook design creativity in favor of feasibility and thus creative designs may be left behind prematurely without much consideration [9-11]. While design creativity evaluation methods do exist [12, 13], they are often linked to methods used in decision-making tools. These methods are generally repeatable and include assessments for design novelty and feasibility, but struggle to establish the “degree of novelty” [14]. Therefore, research that explores and investigates the utility of these design creativity evaluation tools and the development of new streamlined tools can add to our understanding of design creativity and allows us to develop design selection tools that consider both novelty and feasibility effectively.

1.1 Background & Motivation

Designing with a focus on creativity helps to produce designs that are not just technologically advanced but meaningful designs that upset the status quo of stagnant thinking [15]. While creativity is often in high demand early on in design generation, however, the push for it dissipates throughout the design process and creativity is often overlooked during the selection process due to ill-conceived perceptions that creative designs are not feasible or realistic [11]. These notions intensify the importance of defining design creativity and developing tools and methodologies that can quantify it.

In engineering design, creativity is often defined by two parameters: 1) design novelty, which is the uniqueness of the idea, and 2) the idea's utility and design feasibility, which is how well the ideas meet the design goals [16-18]. While design creativity involves both of these factors, the methods currently employed by designers to evaluate design creativity rely on modifications or expansions to tools and methods originally designed for decision making or concept selection that largely ignore design novelty [19, 20]. This is problematic for practicing engineers and students who are, by default, biased against novelty in favor of feasibility during design selection sessions. In fact, researchers have found that it often occurs that novices as well as expert raters have a tendency to get fixated on the intricacies of a product's feasibility and ultimately discard ideas without consideration for novel components of the product [21-23]. Therefore, it becomes essential to encourage creativity throughout the development process.

While not developed for use in engineering industry, there have been several methods developed to rate creativity in engineering design research. For example, Shah, Vargas-Hernandez, and Smith [24] create metrics for evaluating design quality, novelty, feasibility and variety based on feature-tree analysis in order to determine the relative creativity of ideas [25]. Another popular method, Sarkar's SAPPhire, evaluates designs based on the relative uniqueness of the actions of the product, the features that compose the product and the base level quality of the product, and can be mapped to a Function Behavior Structure (FBS) model, which has used in artificial intelligence research [14, 26]. While effective for rating ideas in a research setting, one of the main drawbacks of these methods for widespread implementation is novelty is often measured with multiple highly trained raters who rate ideas based on specifically designed questionnaires [25, 27]. The trained raters are generally Ph.D. students, professors or professional designers with years of experience [28, 29]. For each unique problem set, new rater guidebooks need to be developed, hiring raters can be expensive and training can be time consuming especially for design projects in the classroom with students or in industry [30]. For this reason, creativity evaluation methods are rarely used outside of academic research [31]. Therefore, while these methods provide a means to quantify creativity, their time intensive nature and specificity to each design problem make them difficult to implement in industry and classroom practices.

Because of these limitations, researchers from cognitive engineering, computer science and usability engineering have begun to explore more divergent approaches to understanding design creativity and methods to measure it. Studies in cognitive engineering have helped to understand the role of bias and emotional response within design creativity and have provided insights into opportunities to influence their effects in creativity evaluation tools [32-34]. Other studies in computer science have developed frameworks for computational design creativity tools to aid in the design process through the use of artificial intelligence and natural language processing [35, 36]. Research in product usability and affective engineering has developed several methods to extract user desirability preferences through word selections [37] and quantitative feedback on emotional affinities being conveyed by product designs [38]. These prior studies have pushed our understanding of design creativity by illustrating potentially novel approaches to evaluating creativity, but they have not been empirically explored or validated for design practice.

While divergent design creativity tools are being explored in areas of computer science, and concept selection tools have been available for use in engineering design, they either lack significant validation or have a tendency to overlook creativity for the sake of feasibility [35, 39]. In light of this, creativity assessments have emerged for products and groups of products within academia [19]. However, these metrics and tools can be too time consuming, costly and unreliable when implemented by student engineers and practitioners in industry [28, 31]. With the push for more creativity in the engineering community, creativity assessment tools have an opportunity to act as catalysts in this movement by giving designers the ability to quantify creativity during the concept selection process.

1.2 Objective

The objective of this thesis is to investigate the development and utility of a new creativity assessment tool that is repeatable, considers the role of human bias and is easily implemented in both academic and industrial settings. This thesis is composed of two conference papers, and two journal articles that collectively guide the overarching research goals. Specifically, the goals are as follows:

Goal 1: Explore the utility of word selections and word relatedness as a means to evaluate design novelty and feasibility as they compare to prior creativity assessment methods. Prior studies have shown that word selections can be used to obtain feedback quickly and effectively from users and customers regarding product desirability and emotional impact [37, 40]. However, no studies to date have explored the utility of word selections in creativity evaluations. Therefore, the first goal of this thesis is to investigate how word selections can be used as an accessible and comparable means for quantifying design concept creativity.

Goal 2: Investigate the impact of experience bias on creativity evaluation. Researchers in cognitive science and engineering design have shown mixed findings on the impact of expert and novice raters in the creativity evaluation process [21, 41], leaving it unclear if, and when, novice raters can be used as a proxy for experts during the creativity evaluation process or if expert raters must also be used. Therefore, the second goal of this thesis is to explore the impact of experience bias on creativity assessment in order to determine if novices can be used as a proxy for expert ratings during the concept evaluation process.

Goal 3: *Develop an understanding of the effects of semantic creativity assessment tools in the engineering classroom*. While research in engineering education and product design has studied peer evaluations and self-evaluations [42-44], little research has been conducted that compares subjective self-evaluations of designs with the evaluations obtained from creativity assessment tools in an engineering classroom. Therefore, the final goal of this thesis is to develop an understanding of the difference between subjective and objective creativity assessment tools in order to design recommendations for design creativity assessment tools that can be implemented in the engineering classroom.

1.3 Summary of Thesis Papers

In order to investigate these goals, three manuscripts were developed, and are presented in the following chapters of this thesis. Specifically, the first manuscript published in the proceedings from the *ASME 2014 International Design & Engineering Technical Conferences*

[45], found in Chapter 2, was developed to investigate the utility of word selections and semantic similarity as a method for evaluating design novelty and feasibility. The second manuscript of this thesis submitted to the *ASME Journal of Mechanical Design,* found in Chapter 3, was developed to explore the impact of experience biases on current methods of concept evaluation, including our recently developed semantic creativity assessment tool. Finally, Chapter 4 of this manuscript was developed to explore the impact of the newly developed creativity assessment tool in the engineering classroom. The work of this chapter is to be submitted to the *ASME Journal of Mechanical Design*. These papers culminate in design recommendations described in Chapter 5 for a web-based design creativity evaluation tool that contributes to the efforts to support creativity throughout the design process.

1.3.1 Paper I – A Novel Method for Providing Global Assessments of Design Concepts Using Single-word Adjectives and Semantic Similarity

Published in: ASME International Design Engineering Technical Conference on Design Theory and Methodology, Buffalo, NY, August 17–20, 2014

Engineering design idea-generation sessions often result in dozens, if not hundreds, of ideas. These ideas must be quickly evaluated and filtered in order to select a few candidate concepts to move forward in the design process. While creativity is often stressed in the conceptual phases of design, it receives little attention in these later phases – particularly during concept selection. This is largely because there are no methods for quickly rating or identifying worthwhile creative concepts during this process. Therefore, the purpose of this study was to develop and test a novel method for evaluating the creativity and feasibility of design concepts and comparing this method to gold standards in our field. The semantic creativity assessment tool (TASC) method employed in this paper uses word selections and semantic similarity to quickly and effectively evaluate candidate concepts for their creativity and feasibility. This method requires little knowledge of the rating process by the evaluator. We tested this method with ten engineering designers and three different design tasks. Our results revealed that TASC ratings can be used as a proxy for measuring design concepts, but there are modifications that could enhance its utility. This work contributes to our understanding of how to evaluate creativity *after* idea generation and provides a framework for further research in this field.

1.3.2 Paper II – The impact of Experience Bias on Concept Creativity Evaluations

Submitted for consideration to the *Journal of Mechanical Design*, December 2014

Product design teams are generally guided by engineers with years of experience due to their ability to rely on the successes and failures of previous designs to make more encompassing decisions. Although these individuals afford us the ability to quickly pick out optimal designs rather subjectively and quickly, they can also be impacted by heuristics that may not apply to the current design problem. In following study, we sought to explore the impact of experience biases on current methods of concept evaluation including our TASC method. Our results support the use of novice design engineers for the evaluation of early design creativity as a proxy for expert experience. This work contributes to our understanding of creativity evaluation and provides support for our TASC method and novice perception as evaluation methods.

1.3.2 Paper III –Design Creativity Assessment Metrics in Engineering Education

Submitted to for consideration to the *Journal of Mechanical Design*, December 2014

In design research, creativity assessment methods have been studied to obtain quantitative measurements of design novelty and feasibility for use in the concept selection process. However, little research exists that studies the application and implementation of these tools by engineering students on grade-dependent class projects. In this study, teams of undergraduate engineering design students evaluated their own early product sketches using their individual judgment, team judgment and our TASC adjective selection method. The resulting evaluations were compared and contrasted with evaluations obtained from the widely adopted SVS method used in academia. For team evaluation, our results showed significant positive relationships between the team judgments and our TASC evaluations of creativity ($p < 0.001$), but no significant correlation with SVS evaluations. For individuals, our results showed significant positive relationships between individual judgments of creativity and both TASC ($p < 0.001$) and SVS ($p < 0.05$) evaluations. These findings demonstrate that our TASC adjective selection method of evaluating design creativity is tapping into similar constructs of creativity as the design teams, and also indicate that the SVS method does not appear to be evaluating creativity as perceived by engineering design students. The results from this study can be used to enhance our understanding of design creativity, and to help support creativity in engineering design education.

1.3.3 Summary of Papers

These research papers provide an exploration of the impact of creativity evaluation methods, and their integration into engineering design education, and professional practice. Their findings, and implications contribute to the overall goals of this thesis, and provide a basis for developing new, and more accessible creativity assessment tools for wide usage. The following three chapters (2-4) present these papers in manuscript form and chapter 5 is used to summarize the main contributions of these articles and their limitations.

Chapter 2

A Novel Method for Providing Global Assessments of Design Concepts Using Single-word Adjectives and Semantic Similarity

Published in: ASME International Design Engineering Technical Conference on Design Theory and Methodology, Buffalo, NY, August 17–20, 2014

At the very beginning of the design process, designers and their teams spend a significant amount of time defining the problem and then generating initial designs and ideas to explore. These design idea-generation sessions often result in dozens, if not hundreds, of ideas. These ideas must be quickly evaluated and filtered in order to select a few candidate concepts to move forward in the design process. While creativity is often stressed in the conceptual phases of design, it receives little attention in these later phases – particularly during concept selection. This is largely because there are no methods for quickly rating or identifying worthwhile creative concepts during this process. Therefore, the purpose of Chapter 2 of this thesis was to develop and test a novel method for evaluating the creativity and feasibility of design concepts and comparing this method to gold standards in our field. This chapter introduces the TASC creativity assessment method developed as part of this thesis and an empirical study with ten engineering designers. The results of this study contributes to our understanding of how to evaluate creativity *after* idea generation and provides a framework for further research in this field.

2.1 Introduction

Engineering design research has long since devoted attention and resources to developing tools and methods for supporting creativity during idea generation [46, 47]. While creativity is emphasized in these early phases of design, it is rarely considered in later stages [48-50]. This is problematic because even if designers develop creative concepts, they may not be selected to move forward in the design process. Identifying the *right* ideas is one of the most elusive components of the design process. Steve Jobs once said, "[innovation] comes from saying 'no' to

1,000 things." However, it's also saying yes to the right things – the creative things [51]. 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 [52].

While not specifically focused on creativity, there has been a wealth of research devoted to developing metrics and methodologies to help designers evaluate engineering design concepts [19, 37, 47, 53]. For example, design feature tree analysis, and its derivatives have been prevalent in current design novelty evaluation tools [19, 52]. Although these methods provide an unbiased method to evaluate design concept novelty, they require a tedious analysis process making them ill fit for use outside of academia [48, 49, 54]. In addition, these methods were developed as a means to compare ideation method effectiveness rather than for aiding designers in evaluating candidate design concepts. Therefore, new methods are needed that allow designers to assess design concept creativity more efficiently to ensure designers more thoughtfully consider creativity after idea generation.

Therefore, the purpose of this chapter is to introduce, and test a novel method for evaluating the absolute creativity of design concepts using adjective selections and semantic similarity. This approach requires the designer to have little understanding of design metric calculations, and negligible time to complete evaluations. This work contributes to our understanding of the utility of new metrics for evaluating creativity, and a focus on creativity *after* idea generation. This research directs us to a more efficient system for evaluating design concepts, and supporting creativity in the selection process.

2.2 Design Concept Creativity Metrics and Evaluation

Although not developed for the purpose of aiding in concept selection, there has been a wealth of research devoted to developing effective methods for evaluating the *relative creativity* of design concepts, or the creativity of an idea relative to the other ideas developed in an idea set [55, 56]. In other words, relative novelty metrics are "essentially a measure of whether the exploration occurred in areas of the design space that are well-travelled or little-travelled" [57]. These methods were developed primarily to evaluate, and compare the performance of idea generation techniques [25], and relate the relative "goodness" of design ideas with the idea generation technique used [58]. Cognitive psychologists, and engineering design researchers have developed and adopted two distinct methods for evaluating the relative creativity of design ideas.

Cognitive psychologists typically measure creativity in terms of the (1) quality (i.e., utility of completeness), (2) originality, (3) elegance, and (4) variety of each designer's ideas [59]. In this method, researchers select ideas generated in the study that are high, medium, and low examples of the study criteria (e.g., high originality, medium originality, low originality, etc.). These ideas are used to provide judges with reference points during the assessment process [60]. Once familiarized with these exemplars, judges provide ratings of the entire idea set using 7-point Likert scales. While these rating techniques have been used extensively in the assessment of creative outcomes, and have produced inter-rater reliability values (ICCs) in the range of 0.80–0.90 [e.g., [61, 62], the method relies heavily on selecting appropriate exemplars for comparison prior to the rating process. In addition, the method is extremely time-consuming to complete and is subject to the cognitive biases and limitations of the raters [19].

On the other hand, engineering design researchers typically rely on functional decomposition to measure the relative creativity of design ideas [58]. In this method, researchers collect all of the ideas generated by all of the participants, identify key attributes (e.g., control mechanism, motion type, etc.), and identify all of the ways each design address each of those attributes. A count is then created for each instance of the solution method; the lower the count the higher the novelty because it means few people solved the problem in the same way. This method is widely adopted in engineering design studies because of its unbiased and repeatable feature tree style approach to evaluating design concept creativity [19, 58]. However, Nelson et al. [25] and Srivathsavai et al. [63] have identified the metric's limitations, such as inaccurate representations and poor inter-rater reliability. In addition, this method requires a new feature tree to be developed for each design task, which is a time-intensive and burdensome process.

While these concept evaluation metrics have progressed understanding in the field of engineering design, there remains a need for a more efficient (time-involved), and holistic method for accessing concept "goodness". The methods discussed above do not consider the creativity associated with ideas that are fundamentally novel with respect to the whole of human history (*historic creativity*) [64]. This is important because, although an idea can be novel compared to others idea generated, it may not be commercially viable if similar concepts already exist on the market. Because of this, researchers have begun to explore a more holistic approach to measuring concept creativity, and technical feasibility through Comparative Creativity Assessment [19]. While this method allows for one universal method for idea assessment, it is still based on the feature-tree approach and the equations by Shah, Vargas-Hernandez, and Smith [58].

Due to the timeliness involved in the current idea assessment techniques, and the fact that they were developed to compare ideation effectiveness, few of these methods are utilized in engineering education or industry. Thus, these methods are not useful for identifying the "goodness" of ideas during the concept selection process. The purpose of this study was to develop, and test a novel holistic method of assessing both concept creativity and quality.

2.2.1 Cognitive Biases and Decision Support during Concept Selection

In addition to considering the current methods for assessing concept creativity, it is important to remember that the concept selection process is inherently a decision-making task with human decision makers. Humans and, by extension, product design teams are naturally biased in their evaluation, and selection of novel concept designs [65]. The experience, and heuristics of each designer ultimately results in the decision to rank one idea as being either better or worse than another. Engineering designers with extensive experience, and an understanding of their inner decision making processes can have an edge over others [66, 67]. Whereas designers with little experience and understanding may follow false, or inaccurate heuristics that result in poor decision-making and the selection of early concepts that fail to thrive. In addition, it is well known that designers have an inherent bias against creative ideas due to the risk associated with uncertainty, and risk associated with novel concepts [68]. For this reason, it is important to acknowledge the role of cognitive biases, and heuristics of engineering designers during concept selection.

Subjective bias is defined as decision making or evaluation based on personal, poorly measurable, and unverifiable data or feelings weighted against objective, unbiased data which impede the ability of individuals to make valid decisions or evaluations [69]. In particular, designers, or any evaluator, can be biased based on the order in which they evaluate the ideas (cue primacy and anchoring effects), or their inability to look at the problem from a perspective other than their own (framing bias) [70]. It is widely known that designers perceive the world around them by taking in what is accessible to them, and then filling in the gaps from estimates based on past experiences [67, 71]. This has many benefits especially with regards to making quick decisions, but also makes it possible for designers to misinterpret the information available. Additionally, this bias towards available information can cause designers to make decisions with a limited view of the possibilities [67, 72]. Single features either present, or absent from early concepts can quickly lead to rejection no matter how creative or feasible the concept may be otherwise. A designer or design team's past experience with certain features, materials, or design problems can also lead to poor design selection. There is a great deal of power in decision making that is based on associative experiences no matter how loosely they connect to the current situation [73]. Unfortunately, the majority of product design teams lack the power, and experience of the Steve Jobs and Elon Musks of the world to choose ideas at will without having to rationalize their bases.

In acknowledgment of the limitations of free-form design decision-making, engineering designers are taught early on in their careers a standard set of methods to evaluate designs. Pugh's evaluation method [74], Marsh's Analytic Hierarchy Process (AHP) method [75], Pahl and Beitz's Utility Theory [76], and Quality Function Deployment (QFD) matrix method [77] are just a few of the methods taught to engineers. While these methods are widely used in academic, and industry practices for evaluating concepts, they often neglect to consider the creativity or uniqueness of each concept during the selection process [78]. 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 [52, 79]), these methods are largely unexplored for their impact on creative concept selection or their ability to aid decision makers in the process. The various implementations, and complexities of AHP, and Pugh methods are also becoming problematic in an increasingly fast paced, and innovation-thirsty world of product development [80-82]. Therefore, new methods are needed that minimize the cognitive biases and limitations associated with both selecting creative concepts, and providing input that allow designers to more thoughtful consider creativity in their decision-making process.

2.2.2 Semantic Similarity As an Evaluation Tool

Although not studied in the context of creative concept evaluation, there are proposed decision-making methods in other domains that may be able to be successfully implemented in idea evaluation. Specifically, Benedek and Miner [37] developed a decision system that uses a set of carefully selected words to describe a user's reaction to different product concepts. This method requires individuals to select words from a set of adjectives in order to describe their feelings towards the design concept. Roughly 40% of the words in the set are considered "negative" in order to help evaluators provide more rounded feedback on the concepts and not bias the decision maker. Although this system does not generate a quantitative score for the design concepts, it presents a simple method for obtaining evaluations from decision makers that minimizes the biases associated with asking individuals to "evaluate the concept".

The purpose of the current study is to test the application of this method, combined with the use of natural-language processing, for evaluating concept novelty, and feasibility. Latent Semantic Analysis has been used to analyze design team documents, and to effectively evaluate the performance of design teams [83]. The results of semantic similarity analysis make it possible to extract knowledge from words or groups of words within or between large datasets consisting of text [84]. With the digitalization of (nearly) all human activity available on the web, semantic similarity has been instrumental in applications such as search engine optimization [84], consumer specific marketing tools [85] and data mining [86]. These applications lend themselves to extracting value and making decisions from natural language autonomously.

Combining semantic evaluations with a word selection task may provide an efficient method for analyzing design concepts. 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 [34]. In this study, participants were primed to respond "creatively" to different nouns. Latent semantic analysis was used to determine the semantic distance between the noun and the response. This system was developed as a noninvasive method for assessing the creativity of individuals. These findings lend the possibility that concept designs can be evaluated quickly using word selection, and validity maintained using semantic similarity tools.

One such tool for performing semantic analysis is WordNet::Similarity. This toolkit uses the popular lexical database WordNet as a source of English word attributes for comparison [87, 88]. The six semantic similarity metrics calculated using WordNet::Similarity include three based on least common subsumer (LCS), and the remaining three are based path lengths between words [87]. Of the six measures, the semantic gloss vector measure is most useful for word selection in a concept selection task due to its ability to measure the relatedness of adjectives. The semantic gloss vector is calculated using the context of the word's use in WordNet's sample texts to determine the similarity two words [89]. For example, WordNet links words together similar to a thesaurus by grouping by their relative meaning, and context. So, the words "car", and "automobile" would be closer in similarity than the words "car" and "truck."

While the proposed method 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 and quality metrics.

2.2.3 Research Objectives

The purpose of this research is to introduce, and test our semantic concept assessment tool (TASC) for evaluating the creativity of design concepts through the selection of single-word adjectives, and the calculation of semantic similarities. This method creates two absolute measurements of novelty and feasibility, which are two key components of creativity in engineering design [90]. This method was created without specificity to the design task, which allows it to be universally applied across design problems. TASC is intended for use by design teams for quickly evaluating large sets of early concepts for their novelty and feasibility during the concept selection process. In order to test our approach, we compared TASC to common creativity assessment tools. Specifically, our study was focused on answering the following questions:

1. Does the TASC-innovation ratings of absolute concept novelty relate to commonly used relative novelty measures [58] or perceived novelty assessments? We hypothesize that there will be a weak relationship between the TASC-innovation rating and the relative Shah, Vargas-Hernandez, and Smith [58] novelty measure because of the differences

between absolute versus relative creativity [55]. In addition, we hypothesize that the TASC-innovation ratings will relate more closely to the perceived novelty metrics because they're based on human judgments [65, 91].

2. Does the TASC-feasibility rating of absolute concept feasibility relate to commonly used quality measures [53], or perceived feasibility metrics? We hypothesize that TASC-feasibility measures will be more closely related to the Linsey et al. [53] quality measure that uses questions based on absolute assessments based on prior literature [91, 92].
3. How many raters are needed to most accurately measure novelty, and feasibility using the TASC method or slider-based perceived creativity measures? We hypothesize that the value of additional evaluators will plateau as described in literature on crowdsourcing and online community concept evaluation methods [93, 94] allowing for a recommendation for the number of raters needed.

2.3 Methodology

To answer these research questions, a controlled study was conducted with ten engineering designer professionals and students. This section serves to summarize the methodological approach taken in this study.

2.3.1 Participants

Participants were recruited via emails to engineering design listservs. In total, ten engineering designers (6 females, 4 males) between the ages of 19, and 30 (mean of 22.2) voluntarily participated in the study (there was no remuneration for participation). Seven of participants were undergraduate engineering students, and three participants were graduate engineering students, or practicing designers.

2.3.2 Procedure

At the start of the study, the purpose and procedure of the study were discussed and any questions were answered. Next, implied consent (IRB) was attained. Participants then completed an 81-question online survey that was broken into two parts. Part 1 of the survey (27 questions) required participants to select adjectives that described each of the design concepts presented while Part 2 of the survey (54 questions) required participants to subjectively rate the same concepts for both feasibility, and novelty on a 100-point sliding scale. There were 27 concepts from three design tasks presented during the study that were presented in random order to each participant in order to control for ordering effects. Details of the survey are provided next.

Part 1: Word Selection Questionnaire (WSQ)

In Part 1 of the survey, participants were presented with 27 design concepts and were asked to select five adjectives from a list of 36 words that best describe each concept presented, see Table 2-1 for word list. The word list used in this study was derived from prior research on the Microsoft Desirability Toolkit (MSDT) [37, 95]. These prior studies tested the utility of word selections for measuring desirability of design alternatives in usability studies. Originally, the word list developed by Microsoft consisted of 118 adjectives, but it was later pared down to 55 words through analysis, and testing in three field studies [37].

Once the 55 words were selected from the MSDT, the words were analyzed by the authors for their semantic similarity, or the likeness of each word's meaning, to the words innovative and feasible using the WordNet::Similarity software tool, see section 2.3.3 below for more detail [87]. This was performed to create two numeric indexes of weights for each adjective, see Table 2-1 below. Feasibility, and innovative were selected for the comparison terms because design creativity is often described as ideas that are both novel and technically feasible [52, 58]. However, novel could not be used in the WordNet database due to its association with both the word creative (novel), and book (a novel). The 55 words were then reduced to a set of 36 words after balancing the word list for equal portions of high, and low innovation, and feasibility in an effort to minimize participant bias [37].

Once the survey was developed, design ideas were selected to test our method. The 27 design ideas used in the current study were taken from three prior research studies conducted by the authors. The design tasks in these studies included: “Design a novel milk frother” [96], “Design a novel power mechanism for an electric toothbrush” [97] and “Design a device that minimizes accidents on campus from walking and texting or walking and listening to an MP3 player” [98], see Figure 2-1. In each of these prior studies, the design ideas developed were analyzed for their novelty, and feasibility using Shah, Vargas-Hernandez, and Smith’s [58] novelty metric, and Linsey et al.’s [53] quality metric, and was selected to represent all levels of high, medium, and low feasibility and novelty, see metrics definition below.

Table 2-1: List of Adjectives to Describe Concept Designs with Novelty and Feasibility Coefficients (*N*, *F*) From Wordnet::Similarity

Clear (0.13, 0.11)	Useful (0.10, 0.08)
Compatible (0.11, 0.07)	Expected (0.11, 0.117)
Complex (0.14, 0.12)	Exciting (0.14, 0.1205)
Innovative (1.00, 0.08)	Irrelevant (0.09, 0.11)
Reliable (0.16, 0.08)	Creative (0.13, 0.11)
Busy (0.12, 0.10)	Ordinary (0.19, 0.11)
Clean (0.104, 0.11)	Difficult (0.16, 0.15)
Relevant (0.08, 0.13)	Advanced (1.00, 0.10)
Usable (0.06, 0.05)	Ineffective (0.11, 0.07)
Connected (0.10, 0.10)	Inviting (0.34, 0.10)
Fragile (0.07, 0.06)	Convenient (0.07, 0.09)
Confusing (0.107, 0.11)	Satisfying (0.10, 0.10)
Efficient (0.11, 0.11)	Accessible (0.33, 0.11)
Undesirable (0.08, 0.08)	Comprehensive (0.16, 0.13)
Fun (0.08, 0.06)	Helpful (0.22, 0.14)
Familiar (0.15, 0.08)	Unconventional (0.38, 0.09)
Predictable (0.10, 0.09)	Effective (0.11, 0.11)
Inconsistent (0.09, 0.11)	Powerful (0.15, 0.12)

Three problems were selected for the current study in order to test the utility of our TASC method for analyzing novelty, and feasibility irrespective of the design task. This is important because many popular methods for analyzing design tasks (such as the aforementioned methods) require the development of a feature tree for the specific design problem, which is a time-intensive process [19, 48, 49, 54].

2.3.3 Metrics

The 27 design concepts were rated earlier using the methods of Shah, Vargas-Hernandez, and Smith [58] and Linsey et al. [53].

Design Novelty

Shah, Vargas-Hernandez, and Smith [58] defined novelty as the "measure of how unusual or unexpected an idea is compared to other ideas." As can be inferred from this definition and subsequent metric, Shah, Vargas-Hernandez, and Smith compares novelty relative to the solution set explored [58]. This means that the novelty is not evaluated based on other products that exist on the market or other approaches to solving the problem outside of the generated ideas.

In order to assess novelty using this method, the novelty of each feature was first calculated (see [58] for in-depth discussion). Feature novelty is the novelty of each feature, i, as it compares to all other features addressed by all the generated designs. Feature novelty, f_i, varies from 0 to 1, with 1 indicating that the feature is very novel compared to other features. The method of computing f is

$$f_i = \frac{T-C_i}{T}, \qquad (2\text{-}1)$$

where T is the total number of designs generated for each category, and C is the total number of designs that were rated as being addressed by the design. The novelty of each design, j, is then determined by the combined effect of the Feature Novelty, f_i, of all the features that the design addresses. Because D is computed for all the features, the novelty per design is computed as a percentage out of the total possible design novelty. The method of computing D is

$$D_j = \frac{\Sigma f_k}{\Sigma f_i}, \qquad (2\text{-}2)$$

where f_k is the feature novelty of a feature that was different from the original design, and f_i is the feature novelty of a feature that was addressed in the generated idea.

The 27 design ideas used in the current study were analyzed using this method prior to the study and each of the ideas feature were compared to the entire set of ideas in the respective studies (not just the nine selected for the current study). The nine ideas for each of the three tasks represented three low, medium and high novelty metrics.

Table 2-2: Sample Raw Semantic Index from Wordnet::Similarity

Word	Novelty	Feasibility
Accessible	0.3374	0.1106
Advanced	1.0000	0.1081
Busy	0.1283	0.1059
Clean	0.1040	0.1108
Clear	0.1334	0.1152

Design Quality

After determining the novelty of an idea, it is also important to evaluate the quality of the idea. Quality measures a concept's feasibility or ability [53] to meet design needs [53, 96]. The 27 design concepts described in the previous sections were analyzed as part of prior studies using Linsey et al.'s [53] method. The values obtained from this analysis were calculated by having evaluators answer three questions, "Does it complete the task," "Is it technically feasible to execute," and "Is it technically easy to execute?" Quality is thus evaluated on a 3-point scale, and results in a score between 0, and 1 with 1 being the maximum absolute quality rating [96]. An example of the results from this evaluation can be viewed in Figure 2-1.

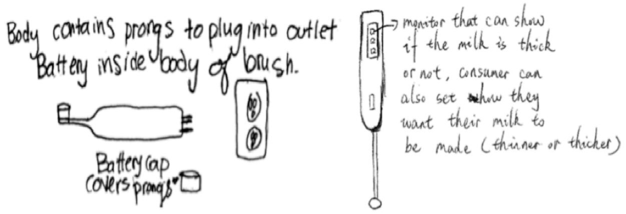

Body contains prongs to plug into outlet
Battery inside body of brush.
Battery cap covers prongs
monitor that can show if the milk is thick or not, consumer can also set how they want their milk to be made (thinner or thicker)

Figure 2-1: Milk Frother (N: 0.69, Q: 1.00) & Electric (N: 0.72, Q: 1.00) Toothbrush Design Idea Sample

TASC-innovation and -feasibility

Evaluation of absolute concept creativity enables the ability to compare design concepts despite the origins of their conception. Relative novelty and feasibility evaluations, as the name would imply, reduce the concept selection space to include only concepts based on similar features or design problems [55, 91]. On the other hand, our method employs a more historic creativity approach, and effort to identify the ideas that are fundamentally novel with respect to the whole of human history (*historic creativity*) [64]. Since creativity is often described as ideas that are both novel and technically feasible, feasible and innovative were selected for the comparison terms for semantic novelty and feasibility scores [52, 58].

TASC-innovation

TASC-innovation scores were calculated in this study according to the semantic distance between each of the 36 adjectives used in the Word Selection Questionnaire (WSQ) and the word innovative using WordNet::Similarity [87, 99]. The semantic gloss vector measure calculated within WordNet::Similarity was selected to measure the semantic distance in this study due to its ability to measure adjective similarity [87] and its leverage of natural language processing technology [34, 37, 84]. Detailed descriptions of how Wordnet::Similarity functions can be found in [87]. A sample of the index used to evaluate semantic novelty can be found in Table 2-2.

Once the weight for each word was determined using Wordnet::Similarity, novelty ratings for each idea were calculated by adding the novelty weights for each of the five words chosen by each participant for each design idea. The method of computing novelty C_{ij} is

$$C_{ij} = \sum_{n=1}^{5} s_n, \qquad (2\text{-}3)$$

where S_n is the semantic value of word n selected by a participant, and C_{ij} is the novelty rating for each i (design problem) and j (design concept).

After the study was complete and the word score for each idea was calculated, the ratings were normalized between 0 (meaning low novelty) and 1 (meaning high novelty). The method of computing TASC-innovation R is

$$R = \frac{C_{ij}}{C_{\max}}, \qquad (2\text{-}4)$$

where $C_{\max}$ is the maximum calculated value of C_{ij} from all the concepts being evaluated.

TASC-feasibility

The TASC-feasibility metric was calculated in this study following the general methodology outlined in the previous section. When evaluating concept designs for feasibility, an index of semantic distance values between the word feasible, and the 36 adjectives on the Word Selection Questionnaire (WSQ) was developed. Technical feasibility is a key factor in design creativity. We again used the semantic gloss vector metric from WordNet::Similarity, see Table 2-2 for a sample index. TASC-feasibility ratings for each of the 27 design concepts for each participant were calculated by adding the feasibility word weights for each of the five words selected by each participant for each design idea, i.e.,

$$D_{ij} = \sum_{n=1}^{5} w_n, \qquad (2\text{-}5)$$

where w_n is the semantic value of word n selected by a participant, and D_{ij} is the semantic value for each i (*design problem*), and j (*design concept*). The method of computing TASC-feasibility R is

$$R = \frac{D_{ij}}{D_{\max}}, \qquad (2\text{-}6)$$

where $D_{\max}$ is the maximum calculated value of D_{ij} from all the concepts being evaluated. TASC-feasibility ratings closer to 0 represented concepts of less feasibility while concepts closer to 1 were considered to have greater technical feasibility.

Part 1: Perceived Novelty & Feasibility

In order to provide an additional metric for comparison with our absolute creativity metric, perceived concept creativity was evaluated. Teams in industry consistently use perceived-creativity assessments to make decisions for their organizations [65]. One-hundred-point evaluation systems are utilized throughout the fields of psychology, education and business to obtain participant feedback [100, 101]. Due to the effectiveness of the 100-point evaluation system, it was used as a subjective evaluation tool within this study.

Specifically, novelty, and feasibility were assessed using two sliding scales – one for feasibility, and one for novelty. This occurred in Part 2 of the study survey. Participants were randomly shown the 27 design concept sketches previously evaluated in Part 1, and asked move the sliders between 0, and 100. Slider positions closer to 0 represented lower ratings of perceived feasibility, and novelty while slider positions closer to 100 represented higher levels of novelty and feasibility.

The nine design ideas selected from each of the three design tasks were selected to represent all combinations of low, medium, and high novelty, and quality (e.g., high quality and low novelty, or medium quality, and high novelty).

Part 2: Perceived creativity ratings

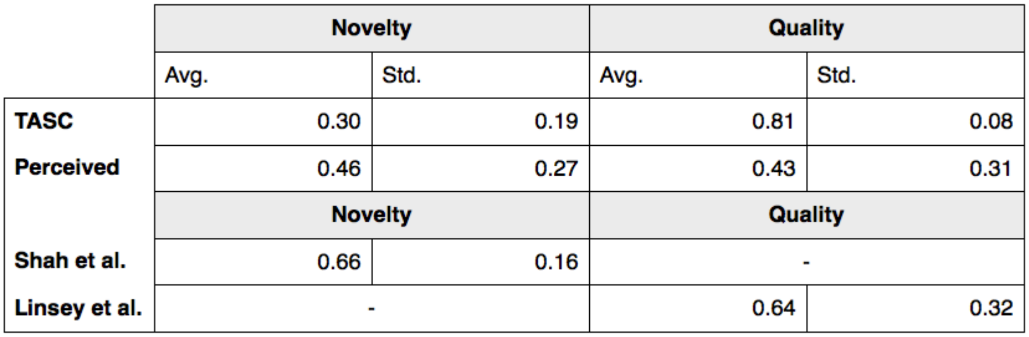
The second part of the survey required participants to provide two ratings for each of the 27 design concepts, presented in random order to each participant. The first rating required participants to evaluate each concept's novelty on a sliding scale from 0–100, where 100 was most novel, and 0 was least novel. Similarly, the second rating required participants to evaluate the perceived feasibility of the design, with 100 representing most feasible, and 0 representing least feasible. This was conducted to compare perceived feasibility and novelty to the other methods of evaluation.

2.3.4 Data Analysis

In order to test our hypotheses, bivariate correlations were computed with the variables of TASC ratings (feasibility and innovation), perceived ratings (feasibility and novelty), and the absolute novelty and quality ratings. SPSS v.22 was used to analyze the findings.

2.4 Results

The following sections present the results of the analysis in relation to our research hypotheses. In Table 2-3, a comparison of average ratings for each metric is shown.



	Novelty		Quality	
	Avg.	Std.	Avg.	Std.
TASC	0.30	0.19	0.81	0.08
Perceived	0.46	0.27	0.43	0.31
	Novelty		**Quality**	
Shah et al.	0.66	0.16	-	
Linsey et al.	-		0.64	0.32

Table 2-3: Comparison of average metric ratings

Relationship between concept novelty ratings

Our first research question sought to identify if there was a relationship between the TASC-innovation score, a commonly used relative engineering design novelty measure, and the perceived novelty score. Our hypothesis was that there would be a weak relationship between the TASC-innovation ratings and the relative novelty measure. We also hypothesized that the TASC-innovation ratings would relate more closely with perceived novelty metrics. Our Pearson Correlation test results revealed a significant relationship between the perceived novelty, and the TASC-innovation score (r = 0.60, p < 0.01). There was a modest correlation between the relative novelty, and the TASC-innovation score (r = 0.33, p < 0.1). There were no other significant relationships. This finding shows that the TASC-innovation scores, and relative novelty ratings were not fully capturing the same information. This was expected due to the differences between absolute, and relative evaluations.

Relationship between concept feasibility ratings

Our second research question sought to identify if there was a significant relationship between the TASC-feasibility score, the absolute quality measure, and the perceived feasibility score. Our hypothesis was that our TASC-feasibility rating would be more closely related to the absolute quality measure. In order to test this, we averaged the TASC-feasibility, and perceived feasibility scores for each of the 27 ideas using the ratings from the ten participants. We then performed a correlation between these ratings, and the absolute quality measure. Our Pearson Correlation test revealed a significant positive correlation between the TASC-feasibility, and both the absolute quality metric (r = 0.40, p < 0.05), and the perceived feasibility assessment (r = 0.47, p < 0.01). Interestingly, there was also a significant correlation between the absolute quality metric, and the perceived feasibility assessment ($r = 0.63$, $p < 0.001$). This finding refutes our hypothesis.

How many raters are needed for the TASC method?

Our final research question sought to address how many raters are needed to most accurately measure design innovation, and feasibility using the TASC method. Our hypothesis was that the having additional evaluators would increase the accuracy of the system, but that accuracy would plateau. In order to address this, TASC-innovation, and -feasibility scores were computed using an average from two, four, six, eight and ten (all) participants. Pearson Correlations were performed between each of the averaged TASC-innovation, absolute, and perceived novelty scores. The results between TASC-innovation, and the absolute novelty metrics had significance with eight, and ten participant scores at the 0.05 level ($r = 0.33$, $r = 0.326$). The correlation results also showed that the TASC-innovation scores were all significantly correlated with one another at the $p < 0.001$ level ($r > 0.80$). This result shows that, for the TASC-method, regardless of the number of participants included in the aggregate rating, each idea was receiving similar TASC-innovation rating. When comparing the TASC-innovation averages with the relative creativity score, six, eight, and ten raters were significant at the 0.05 level ($r < 0.35$). This is unsurprising since the measures are looking at relative versus absolute creativity. Similar results were found for the perceived novelty ratings: all perceived novelty values were correlated at $p < 0.01$ level ($r > 0.89$). However, the perceived, and absolute creativity metrics did not correlate significantly.

When looking at concept feasibility, our results showed that the averaged perceived-feasibility scores (two, four, six, nine and ten), and the absolute feasibility measurement were all significant at the $p < 0.001$ level ($r > 0.68$). This result means that, regardless of the number of participants included in the averaged ratings, each idea received a similar perceived-feasibility rating to the absolute measure. When testing the correlation between the absolute feasibility metric, and the five aggregate ratings for each of the TASC-feasibility metrics, we found that none of the averaged ratings for the semantic feasibility were significantly correlated.

2.5 Discussion

The main goal of this study was to investigate the use of adjective selection, and semantic similarity to evaluate early concept designs for their absolute creativity. Our main results were as follows:

- TASC-innovation ratings can be used as a proxy for relative measures with seven raters, although the predictive value is relatively low (*r* = 0.35);
- TASC-feasibility is not a significant predictor of design quality, but perceived feasibility can be used as a proxy for absolute feasibility methods; and
- Perceived ratings can be used to evaluate both feasibility, and novelty in one rating scheme.

The implications for these findings are presented next.

Semantic Ratings for Novelty and Feasibility

Our first two hypotheses were concerned with the effectiveness of TASC ratings both as a method for measuring absolute creativity, and individual perception of design creativity. Current design creativity metrics used by academia have taken on the challenge of trying to quantify design creativity [52, 58, 84]. Although these methods have made progress in the overall study of design creativity, their ability to compete with human intuition, and experience for evaluating absolute concept design creativity has been limited. The time, and resources required to use previous methods have also hindered their acceptance [19, 52]. The TASC-innovation, and -feasibility ratings, as presented in this study, broke away from the feature-tree analysis methods used previously, and approached creativity evaluation alternatively using natural-language processing technologies [87]. The results from this study showed that TASC ratings could bring us closer to understanding human intuition for creativity. In addition, this study shows the potential for word selection, and semantic distance being used to analyze concepts within the context of creativity. The relationships found between TASC ratings, and perceived ratings highlight this assessment. Ultimately, understanding creative intuition will bring about new tools, and more effective methods for evaluating concept creativity.

Although our work provides insights into the impact of divergent approaches to evaluating absolute concept creativity, more work is needed to examine the role of natural-language processing, and similar technologies. This is important because human intuition for creativity is ultimately articulated through verbal expression. Research outside of engineering design and within psychology, and neuroscience have shown that creativity of individuals can be measured using word selection [34]. In addition, the adjectives used in this study originally were developed to assess desirability [37]. More work can be done to refine this word list, and to better understand why some words are chosen while others are not. Adjective selection analysis may bring about another level of dimensionality to understanding human intuition for creativity.

Across-problem Evaluation

Previous methods of early concept design evaluation and selection have been restricted to comparing ratings within specific individual problem sets [19, 53, 58]. This is due to the observance of relative creativity. A goal for our study was to explore the possibility of being able to evaluate, and compare the creativity of concepts from different problem sets in terms of absolute creativity. Throughout this study a great deal of effort was made to enable across-problem rating evaluation. To support these efforts, three different design problems were analyzed, and compared throughout this study. The three design problems were as follows: “Design a novel milk frother” [96], “Design a novel power mechanism for an electric toothbrush” [97] and “Design a device that minimizes accidents on campus from walking, and texting or walking and listening to an MP3 player” [98].

The TASC rating method used in this study was able to be administered without any regard to the design ideas and design problem being evaluated. This contrasts with other approaches that require a feature tree to be developed *a priori*. Using TASC ratings, it was possible to draw comparisons between concepts from three different problem sets using the same evaluation parameters. The word selection, and semantic rating method shown in this study remove previous obstacles, and approaches creativity evaluation in a simplistic yet sophisticated way.

With this being said, our work hopes to provide a base for future development with regards to evaluating creativity across different design problems. Additional work is needed to examine and compare how humans are able to compare ideas without common features. Framing the design problem is an essential part of the design process [102]. As such, understanding problem differences may provide a means to appropriately frame design problems early on before concept generation.

Developing New Metrics for Industry and Education

Creativity motivates design, and brings forth advances in both industry, and academia [19, 46, 47, 52]. For this reason, our study explored the development of a creativity metric that can be used by both sectors. Previous methods of creativity evaluation, although supported in academia, have failed to be integrated into industry practices [84]. This can be attributed to the amount of time, resources, and complex methodologies required for their use. Through the use of word selection and semantic similarities, this study was able to simplify the evaluation process, and reduce overall resources (time, money, etc.) while obtaining practical results. Both industry, and academia can reap the rewards of using word selection, and semantic ratings. This study brings forward the importance of understanding the needs of both industry, and academia in design research [103].

Another factor that can influence the flow of creative ideas during concept selection is the criteria used for evaluation [104]. In fact, many companies have acknowledged that they have problems establishing clear criteria for concept selection [105], and thus often perform poorly at selecting their own most promising ideas [106]. Creativity metrics may serve to overcome this obstacle, as they have been developed in engineering design to aid in the rating of the feasibility, and novelty of ideas generated during concept generation [58]. While these metrics may prove useful for evaluating design concepts, they have been almost exclusively used as a means to compare the effectiveness of idea-generation methods. However, a recent study explored the utility of alterations of these metrics and developed a new metric, the Comparative Creativity Assessment (CCA) metric, for use during concept evaluation [19]. While that research provides promise for developing, and using creativity metrics during evaluation, the utility of the metric, its generalizability across problem structure, and domain, and its impact on creative concept selection have yet to be explored. Therefore, it is unclear how novelty metrics can be used to aid in the movement of creative ideas through the concept selection process.

While the current study adds to our understanding of creativity evaluation, it was not evaluated specifically in regards to either industry or academic practices. Therefore, future work is needed that implements, and tests our method in these settings. It will be important in future work to investigate the role of designer experience while using our method [94]. However, the results of this study are promising, and overcome some of the challenges of existing strategies.

2.6 Conclusion

The results of this study identify aspects of semantic metrics for evaluating concept designs that impact the current understanding of creativity, and concept selection. Specifically, semantic metrics show promise for evaluating concept novelty within reason, while also requiring less time, and fewer resources. The results from this study also provide significant contributions to the engineering design community by exploring new, and divergent metrics to evaluate concept creativity, and to better understand designer intuition for creative designs.

Future studies should examine the role of evaluator populations in semantic metrics, experimental settings involving evaluation durations, and semantic feasibility ratings. Although word selection was used to mitigate bias in the TASC method, there is an opportunity to study the effects of the design team evaluating their own concepts using the TASC method as evaluators. In addition, new adjectives used to describe design concepts should be explored. Additional research is needed to build on and streamline the use of semantic similarity software tools Finally, the role of semantic, and word selection metrics need to be studied in order to better understand design creativity, and to support creative design acceptance.

Chapter 3

The Impact of Experience Bias on Concept Creativity Evaluations

Submitted to the *Journal of Mechanical Design*, December 2014

The following study served to highlight the utility of the TASC assessment tool to obtain expert grade evaluations from novice raters. However, we know that individual evaluators have their own set of preferences, and judgments for what can be considered creative based on their individual differences, and past experiences. While product design teams are generally guided during the concept-selection process by engineers with years of experience, little is known about how expert, and novices compare in their evaluation of creative concepts or how the TASC method can be used as a proxy for expert responses. Therefore, the goal of Chapter 3 of this thesis was to develop, and to explore the impact of experience biases on concept evaluation in subjective ratings through an empirical study with ten expert, and ten novice designers. The results from this study can be used to develop concept creativity, and selection methods that can utilize raters with novice levels of experience to obtain feedback that is comparable to that obtained from experts. This would enable design teams to iterate numerous times on designs with little cost in comparison to hiring, and scheduling expert-level raters.

3.1 Introduction

Innovation is a crucial component of long-term economic success [14]. As such, engineering design research has long since devoted attention and resources to developing tools and methods for supporting creativity during idea generation [46, 47]. While the goal of these methods is to help designers generate a large quantity of effective solutions and explore a larger solution space [24], the creative ideas developed through these methods are often rapidly filtered out during the concept selection process [106]. In other words, while creativity is often emphasized in the early phases of design, it is rarely considered in later stages [48-50]. This is problematic because, even if designers develop creative concepts, they 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 [106], 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 [52].

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: Pahl, and Beitz's Utility Theory [76], Marsh's Analytic Hierarchy Process (AHP) method [75], Pugh's evaluation method [74], Quality Function Deployment (QFD) matrix method [107], and Thurston's fuzzy-set method [108] (see [109] 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 [78].

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 [52, 79]), 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 to help designers evaluate engineering design-concept creativity [19, 37, 47, 53], these methods are rarely used outside of academic purposes due to the time involved to analyze design concepts. 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 quantitative approach for evaluating concept creativity is to rely on independent reviewers' subjective agreement [59]. 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 [35]. Despite the speed behind human perception, however, judgments can be inconsistent, and lack quantitative support [13, 110]. While expert designers are often used to evaluate candidate designs based on their experience, interestingly there has been little research geared at exploring the difference between expert, and novice ratings of concept creativity.

Therefore, the purpose of this chapter is two-fold. First, we seek to identify perceptual differences in concept creativity, novelty, and usefulness [18, 59], between expert, and novice engineering designers across three problem domains. Second, we seek to introduce, and test a novel method for evaluating the absolute creativity of design concepts using adjective selections, and semantic similarity. This approach minimizes human biases, and 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 a focus on creativity after idea generation. This research directs us to a more efficient system for evaluating design concepts, and supporting creativity in the selection process.

3.2 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 [19]. The majority of this research has focused on relative measures of a concept's creativity compared against other ideas in the same generated set [55, 56]. The relative nature of these metrics help to inform the designers about the uniqueness of the ideas within the design space [25] and compare the relative creativity of the concepts generated [111].

In the field of engineering design, relative creativity is often measured by breaking down the design concepts into their unique features [112]. For example, the widely adopted, and gold standard in engineering design, 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) [24]. 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 [25, 111]. 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. This method of decomposition, and feature-tree analysis has become a gold standard in engineering design research due to its unbiased nature, and repeatability [19, 24]. Despite the wide use of this method, however, many limitations have been reported such as low inter-rater reliability leading to extensive rater training, inaccurate representations, and difficulties interpreting multiple metrics simultaneously [63, 113].

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 variety of the concepts [90]. This evaluation begins with the selection of anchor concepts for high, medium, and low creativity within the generated set [60]. With these anchors, judges are trained to evaluate other concepts on a relative basis. Afterward, the actual concepts generated are evaluated using 7-point Likert scales. This method has been used widely to assess creativity with strong inter-rater reliability values in the range of 0.80–0.90 [61, 62]. Despite the widespread adoption of this method in cognitive science, however, it requires careful selection of the anchoring design examples, and can easily be exposed to cognitive biases based on the expertise of the design evaluator [114, 115].

Because of the deficits of existing approaches, researchers have begun exploring alternative methods for evaluating concept creativity through the development of computation design creativity systems (CDC) [30]. CDCs provide an opportunity to leverage computational power, and review large data sets to support creativity evaluations that consider historical creativity. The development of more robust creativity frameworks could be the key to enabling CDC systems. The work of Maher, and Fischer [116] has sought to more appropriately characterize product creativity for use within CDC systems, and for the development of artificial-intelligence (AI) systems to evaluate design creativity. This research judged creativity under the characteristics of novelty, value and surprise with consideration for a blend of both relative and absolute creativity. The work of Gero, and Kannengiesser [36, 117] has also sought to enhance CDC systems in AI 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 [24, 30].

The deficit of current evaluation methods, and the emergence of CDC systems supports the opportunity for future creativity evaluation metrics. However, these systems have not been thoroughly tested outside of experimental research. Therefore, the goal of this research is test the effectiveness of creativity evaluation methods in comparison to human perception of creativity as it relates to product design.

3.2.1 The Role of Experience in Creativity Evaluation

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 do to pattern recognition [41, 118, 119]. Using this, experienced individuals can make evaluations quickly, and rather easily, whereas inexperienced individuals may get bogged down by reviewing all given information. It would follow then that, when solving problems or designing, it is possible for experienced designers to make reasonable decisions based on automatic processing. However, this automated processing may also lead to individuals disregarding important, or subtle information that an inexperienced individual will retain [21]. While there is a general support of the use of expert raters in the cognition literature, it has only been recently that engineering design researchers began to explore the role of expertise on design-concept ratings.

To obtain a base of understanding regarding the influence of expertise during concept evaluations, and selections, it has been important to explore research in other areas such as cognitive science, and decision-making theory. The research performed by Sun et al. [32] showed that experienced designers were more adept at generating numerous designs that were also of high quality in comparison to inexperienced designers. These findings were supported by cognitive workload analysis that showed increased creative process efficiency in experienced designers. However, more recent research by Green et al. [120] has shown that it may be possible to use novice designers to evaluate design creativity with minimal training, and achieve expert-level feedback. In their study, students were used to evaluate the originality of different designs, and compared to prior ratings by experts. Their research suggests that ratings from students with high inter-rater agreements can obtain expert-level evaluations, and that minimal training with examples can also impact the fidelity of originality ratings. The contradiction in these findings also shows that generation and evaluation skillsets may be different in the design domain.

With mixed research on the role of experience in design and design-like processes, it is important to consider the inner problem-solving and decision-making strategies that guide experienced, and inexperienced designers [121, 122]. In this way, we can extract a framework to guide design creativity tools to account for designer experience. A case study using an experienced industrial designer showed how small heuristics were used by the designer to effectively explore the problem space, and develop more creative solutions [123]. While not explicitly studied, these smaller, and quickly formulated decisions might also impact the concept-selection phase in the design process. In another study, experienced designers were shown to describe more efficiently concepts by removing any extraneous words [124]. These results also carried over to expert sketches that contained less detail, and were more organized than those of novices. These findings align with the research previously described in cognition research regarding automatic processing of information due to expertise, and context [125].

As researchers gain more information about designer experience, and the associated heuristics, there is a greater opportunity for design-creativity tools, and processes that help novices in education, and industry to choose designs that are both of quality, and ultimately creative. Within the constructs of our current study, we sought to understand the successes, and limitations of current creativity evaluation methods that seek to reduce bias in concept selection. In this way, improvements, and modifications can be made to strengthen the capabilities of future evaluation tools.

3.2.2 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 [126]. Affective design refers to the process of creatively engaging the customer’s emotions in such a way as to differentiate one design from another design [127]. In order to achieve this, researchers have utilized Kansei and subjective methods to quickly evaluate human perception [128], satisfaction [127] and desirability [37] as a means to develop new, and 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 [129]. This process of obtaining adjectives helps designers to create a model for how customers are interpreting the designs. Adjectives are clustered in this method by how they match a specific design factor [38]. This categorization process is similar to that of the relative measures of creativity involving feature-level analysis [63] 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” [40]. 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 [38, 40, 127]. This method of design analysis has the rigor, and relevance of Shah, Vargas-Hernandez, and Smith’s method [130], 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 [37]. 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 [37]. 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 don’t 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 purpose of this chapter was to explore the use of affective engineering techniques, and compare this method to existing relative methods.

3.3 Research Objectives

Prior work has discussed the role of experience in cognitive science, and psychology 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 metrics, and the efficiency of human perception, see Figure 3-1. However, there has been little research conducted that investigates the impact of experience on the design evaluation methods themselves. 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. Do experts' and novices' perceptions of ideas differ in terms of design 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 [131, 132]. There has also been research identifying the tendency for novices to rely heavily on personal experience to evaluate design feasibility, but they lack the personal, and domain experience of experts [133, 134]. Therefore, we hypothesize that there will be differences among expert and novice perceptions of early phase ideas.
2. How does our TASC method compare to human perception of creativity, and the relative SVS [24] method ? Although relative measures of concept creativity enabled reliable, and repeatable analysis of creativity in engineering design research, they are difficult to implement in industry, and lack the ability for comparisons across multiple problem spaces [30, 135]. Human perception, on the other hand, allows for fast-paced analysis, but at the risk of relying on cognitive biases that can be flawed in their assumptions, and resolutions [54, 136]. 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. Does the TASC method, and SVS method align with expert human perception, and can TASC be used as a proxy for expert ratings? Prior research has been conflicting in deciding if novices can be used to 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 [137, 138], whereas more recent literature finds that it is possible to obtain expert-level ratings from novices [120]. Therefore, 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.

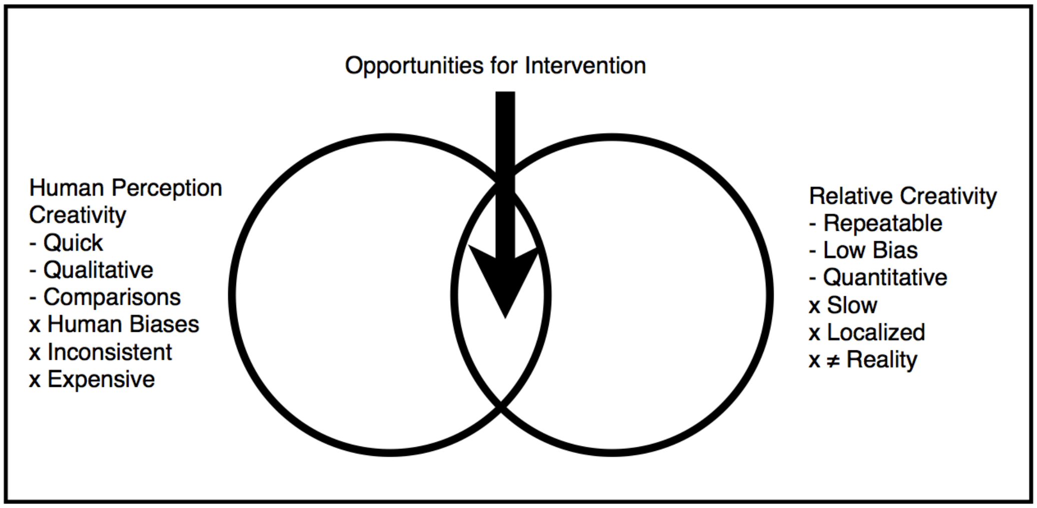


Figure 3-1 Venn diagram comparing design creativity evaluation methods

3.4 Methodology

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

3.4.1 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 (11 novices) while the remaining eleven participants were considered expert engineering designers (11 expert). Ten of the eleven expert participants had engineering design–related advance degrees ranging in areas of focus from human computer interaction to automotive textile product design.

3.4.2 Experimental Procedure

At the beginning of this research study, the procedure, and purpose of the study was presented to the participants and any questions were answered. The participants then completed an 81-question survey that was broken into two parts: (1) the Adjective Selection Questionnaire (ASQ), and (2) the Perceived Creativity Ratings. There were nine concepts from three different design tasks, described below, for a total of 27 concepts. The details of the questionnaire and design concepts tested 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” [96], (2) “Design a novel power mechanism for an electric toothbrush” [139], and (3) “Design a device that minimizes accidents on campus from walking, and texting or walking and listening to an MP3 player” [140]. 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 [30].

The 27 design concepts selected for this study were analyzed for their novelty using the SVS method’s novelty, and quality measures (see [24] for description of this procedure) in prior studies [96, 97, 140]. 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).

Part 1: Adjective Selection Questionnaire (ASQ)

Once the ideas were selected, an Adjective Selection Questionnaire (ASQ) was developed. In this questionnaire, the 27 design concepts were presented to participants one at a time, and participants were asked to 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 [37, 95]. The MSDT contains a list of 55 words that were selected, and tested in three field studies [37]. In the current study, we analyzed these 55 words for their semantic similarity, or relative likeness in meaning [141], 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 [24, 52]. DISCO is an online and downloadable Java class that computes the distributional similarity between words using co-occurrences [99]. 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 [37]. It should be noted that a negative value was assigned to negative/neutral adjectives during the coding process in order to account for bias [37]. DISCO was used in the current study due to its strong correlation with human judgment [99]. The semantic distances calculated during the selection process were also used to create two numeric indices of weights for each adjective, for more detail on semantic weights please see section 3.4.3 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 3-1. 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.

Part 2: Perceived Creativity Ratings

Once participants completed the ASQ the 27 design concepts were again presented to the participants in random order. However, instead of selecting adjectives, the participants were asked 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. Once the perceived ratings were complete, the study was concluded.

Table 3-1 Index of the 36 adjectives for evaluators to choose from including TASC semantic weights used for calculations (Innovative weight, Feasibility weight)

Adjectives and TASC Weights

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)

3.4.3 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 [96, 97, 140]. SVS defines novelty to be “how unusual or unexpected an idea is as compared to other ideas” (p. 117) [24]. In this way, the 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 [14]. The SVS methods generally involve the assessment of feature novelty where the novelty of a feature is relative to the other features in the entire design set. 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 (see [96] for in-depth discussion) [1]. They define quality to be “the feasibility of an idea, and how close it comes to meet the design specifications,” (p. 117) [24]. In the current study, the quality values were calculated by having human 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 [24]. 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 [14]. This combined-creativity score was used to compare, and contrast with other creativity metrics, and between all three design problems.

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. 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

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

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

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

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 [24, 52]. Specifically, the TASC-creativity rating is calculated by taking a direct sum of the TASC-innovative, and TASC-feasible ratings, i.e.,

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

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

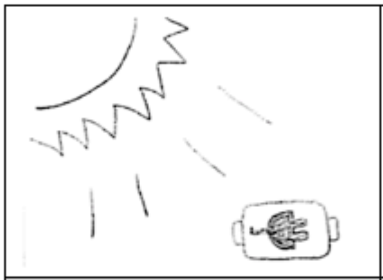
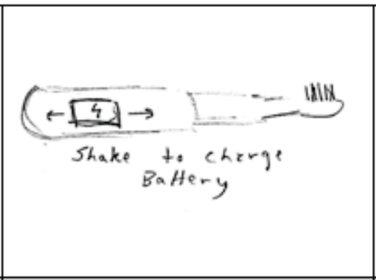
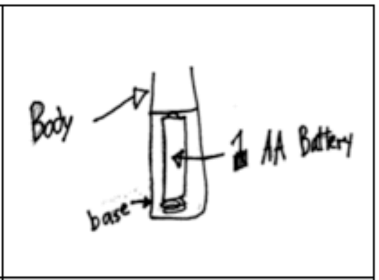
Finally, in order to understand how design engineers perceive novelty, and feasibility subjectively, each of the 27 design concepts was evaluated using 100-point evaluation scales in the second part of the study. This type of evaluation has been used in industry to help teams provide feedback, and make decisions [65]. One hundred-point evaluation systems have also been utilized throughout the fields of psychology, education, and business to obtain feedback [100, 101, 142]. 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.

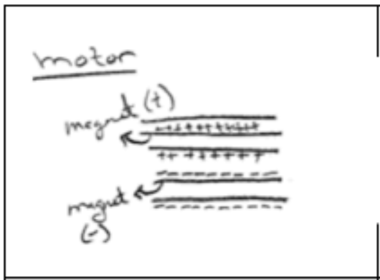
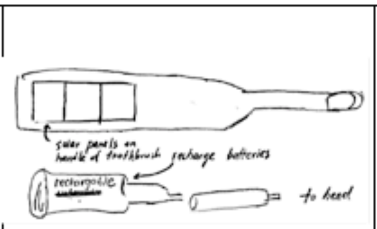
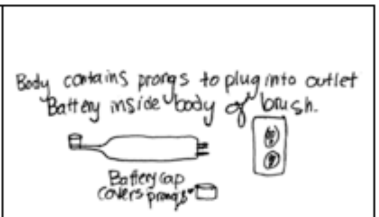
To provide an overall evaluation of perceived creativity, a perceived-creativity composite rating was also calculated by taking a summation of the novelty rating, and the feasibility ratings provided by each participant for each concept that was evaluated. This composite rating was used to compare, and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

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

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 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).

			
	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

		 solar panels on handle of toothbrush recharge batteries rectangular to head	 Body contains prongs to plug into outlet Battery inside body of brush. Battery cap covers prongs
	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

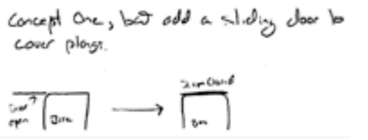
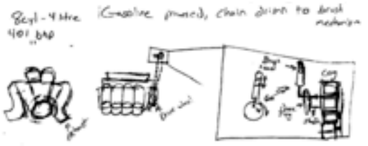
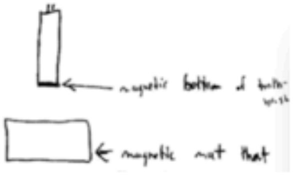
	 Concept One, but add a sliding door to cover plugs. open closed		
	Design 7	**Design 8**	**Design 9**
Perception	(M = 0.22, SD = .22)	(M = 0.57, SD = 0.30)	(M = 0.52, SD = 0.31)
TASC	(M = 0.25, SD = 0.05)	(M = 0.41, SD = 0.22)	(M = 0.43, SD = 0.25)
SVS	0.39	0.70	1

Figure 3-2. Summary comparisons of design evaluations from "toothbrush" design problem.

3.5 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 3-3. The nine "tooth brush" design sketches evaluated in this study are shown in Figure 3-2 to provide an example of the average scores obtained from each evaluation methods. 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" [143]. The following sections present the results of the remainder of our analysis in relation to our research hypotheses.

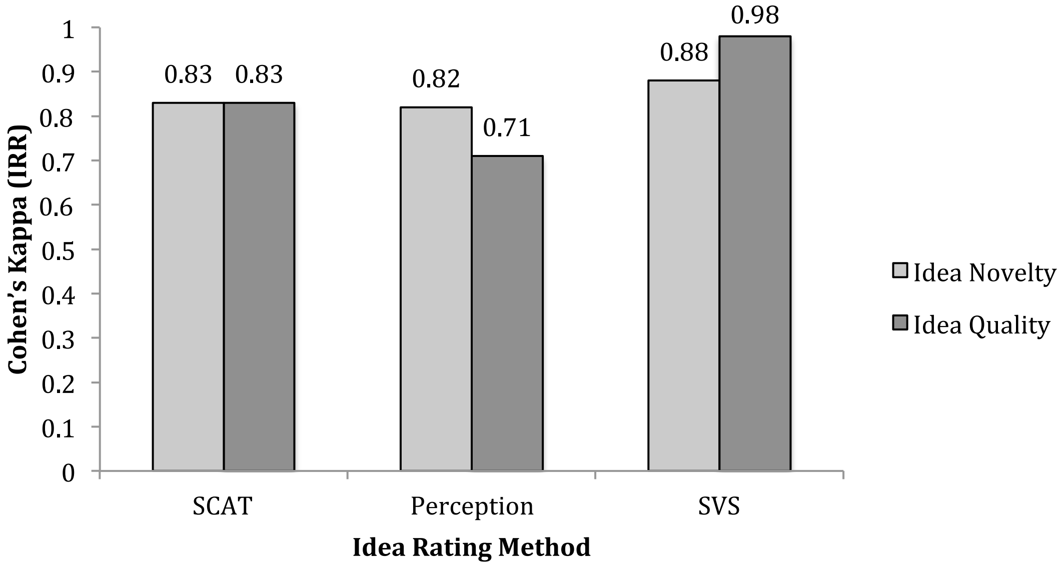


Figure 3-3. The inter-rater reliability (Kappa) for the idea rating methods used in the current study

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. The two-tailed tests of significance indicated that a positive significant relationship between expert, and novice perception of design concept novelty (r_s (9) = 0.741, $p < 0.01$), quality (r_s (9) = 0.749, $p < 0.01$) and creativity ($r_s = 0.861$, $p < 0.01$).

These findings show that aggregate ratings from ten untrained, novice designers can be used as a proxy for expert design ratings. 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 [144]. However, the novice designers in the current study received no training on the design tasks or rating scheme, and our results indicated that only ten 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 [145]. 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 correlations with both measures. As shown in Figure 3-4 there is an area of overlap between the SVS, and human perception methods of evaluating designs that our TASC method has been developed to transcend. 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, a Spearman's Rank correlation was conducted between the novice designers' ratings of idea creativity, and the ratings from the TASC, and SVS methods, see Figure 3-4. The two-tailed tests of significance indicated that there were significant positive relationships between the novices' perception of idea creativity, and the TASC method for the toothbrush (r = 0.98, $p < 0.01$), and the texting (r = 0.88, $p < 0.01$) methods, but not the milk frother (r = 0.39, $p < 0.30$). This finding indicates that the TASC method is tapping into similar constructs of design creativity as the novice perception. On the other hand, there was no significant relationship between the SVS method, and novice perception for any of the design problems tested.

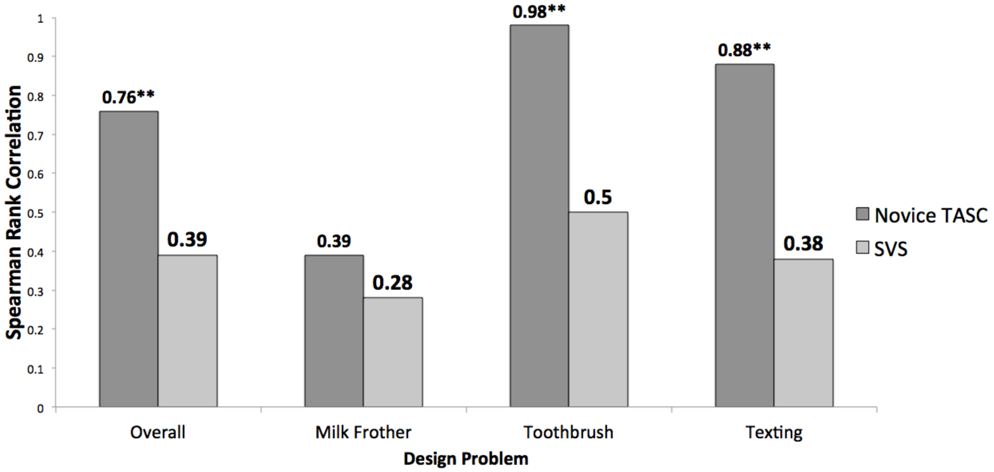


Figure 3-4: A summary of the Spearman's Rank correlations between novice rater's perception, novice TASC and SVS scores of all 27 designs. ** Significant at $p < 0.01$, * $p < 0.05$

Two-tailed Spearman's Rank correlation tests of significance were also conducted between expert designers' perception of creativity, expert TASC scores, and the SVS method, see Figure 3-5. The results showed no significant relationship between the SVS method, and expert perception for the milk frother (r = 0.47, $p < 0.20$), and the toothbrush design problem (r = 0.65, $p < 0.06$). However, there was a significant relationship between the SVS method, and expert perception for the texting problem (r = 0.56, $p < 0.01$). Interestingly a reverse finding was found for the relationship between expert perception, and the TASC method; significant relationships were found for the milk frother (r = 0.75, $p < 0.05$) and the toothbrush (r = 0.80, $p < 0.01$) design problems but not the texting problem (r = 0.60, $p < 0.08$). This may be due to the openness of the texting problem, and the variety of solutions generated for this problem. In these cases, prior works have shown that case-based knowledge, although beneficial in most cases, can conversely cause erroneous conclusions from experts when conditions are not explicitly within the evaluator's perceived domain knowledge [146, 147]. 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[138].

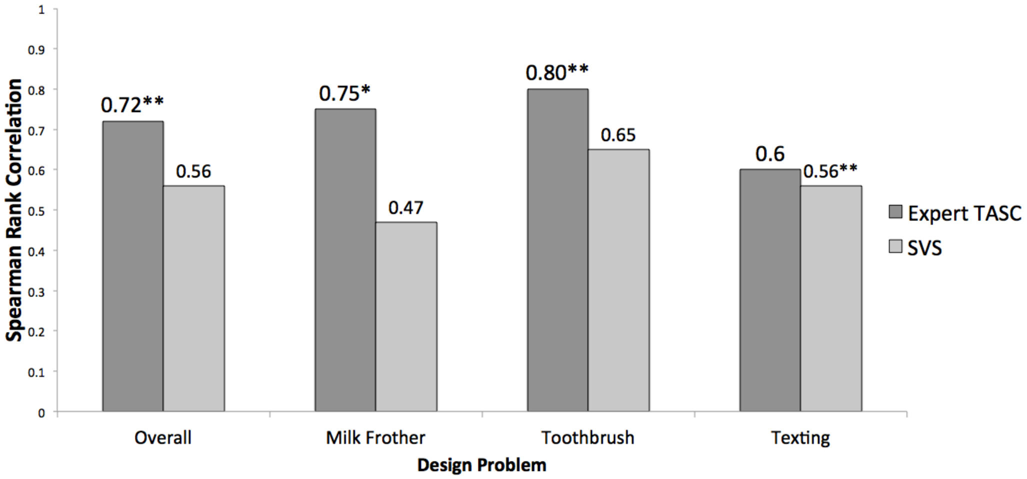


Figure 3-5: A summary of the Spearman's Rank correlations between novice rater's perception, novice TASC and SVS scores of all 27 designs. ** Significant at $p < 0.01$, * $p < 0.05$

These findings support the ability of our TASC method to evaluate design creativity using word selections, and semantic distances. The TASC appears to be tapping into similar constructs of design creativity as hypothesized. This means that the TASC could be used as a proxy for pure expert, and novice perceived evaluations of design creativity. The results from using the SVS method were interesting in that novices had a more significant relationship than experts. This finding could be problematic for the use of SVS methods that may not be tapping into a similar view of creativity as perceived by experts in product design.

The results from these tests continue to support our hypothesis that our TASC method would tap into similar constructs of creativity as human perception. This finding is promising for the development of more absolute, and global measures of design creativity. In this way, the TASC supports the effort to allow the same metric and framework to ultimately evaluate different design problems.

Does the TASC method, and SVS 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, will 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 Spearman's Rank correlations were conducted. Summaries of this analysis, shown in **Figure 3-**6, are provided below. Our results showed a significant positive correlation between expert, and novice perceptions of creativity for each of the design problems, including the "milk frother" (r_s (9) = 0.87, $p < 0.01$), the "toothbrush" (r_s (9) = 0.90, $p < 0.01$), and the "texting" design problem (r_s (9) = 0.82, $p < 0.01$). 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 TASC scores also had a positive significant correlation with expert perception for the toothbrush (r_s(9) = 0.87, $p < 0.01$) and texting problem (r_s(9) = 0.65, $p < 0.05$), but not the milk frother problem (r_s(9) = 0.87, $p < 0.17$). However, the correlation coefficients were not as high for the novice TASC scores as they were for the novice perception scores.

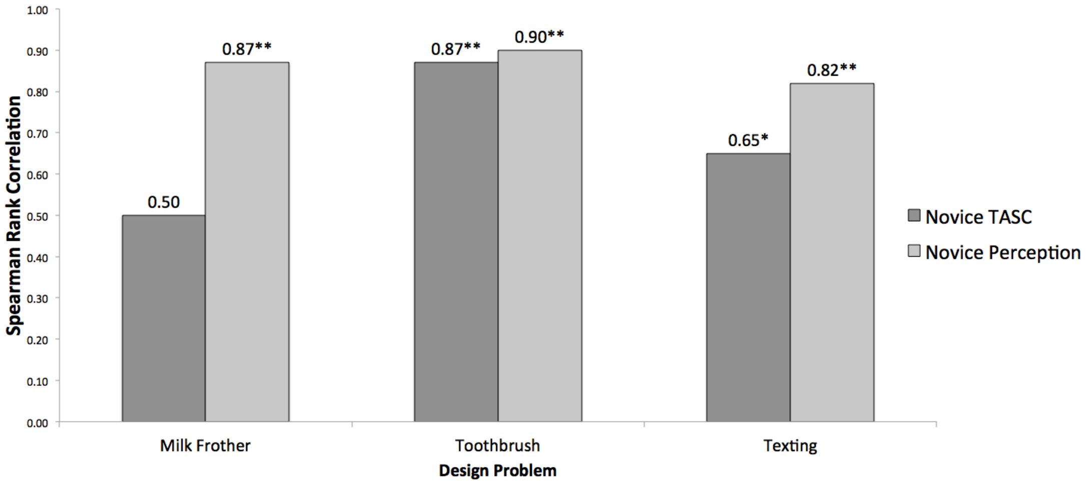


Figure 3-6: A summary of the Spearman's Rank correlations between expert perception and both expert TASC and SVS scores by design problem.

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 ten 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 [40], and the desirability toolkit [37]. Although our TASC method was significant for two of the three design problems, future work will be required to further reduce discrepancies. This could be possible through the development of a crowd-sourced semantic similarity index using Amazon Mechanical Turk [84] 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. In comparison, the SVS method ratings were only significant for the "texting" design problem. 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 [144]. 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 investigate the development and use of our TASC in comparison to human perception and prior creativity metrics, and to further our understanding of how expert, and novice perceptions of creativity relate to other measures. 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 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 of 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, in the context of design concept evaluation and selection, experts and novices are able to reach similar judgments on design novelty, quality, and overall creativity. Despite their varying levels of experience, these two evaluator groups are able to reach similar conclusions about a design's creativity. Our results align with prior research in design expertise and crowd-sourced design that has suggested that novices with minimal training can be used as a proxy for expert feedback [120, 148]. However, there was no training involved in our study and, as our results show, novices were able to yield significant results. Creativity identification in this way is not limited to the expert domain. It also means that training might not be as essential as previously thought.

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 [93, 94], 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 [9, 11]. 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 [37] 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. 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.

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 [99] 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 [48, 120]. Further experimental investigations on this topic can be implemented within our TASC methodology with relative ease.

3.8 Conclusions

The main goal of this study was to investigate the utility of our TASC method, and 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 novices and expert evaluators are able to perceive creativity alike, and that it may be possible to utilize this ability to reduce the costs, and limitations of using purely expert evaluations in the concept evaluation process. However, creativity evaluation methods need further development towards minimizing contextual rater biases for unique problem sets. Lastly, our results showed support for computational design-creativity tools, and their ability 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 for students and industry. Our results are used to provide directions for future research, and provide recommendations for design evaluation and to support creativity throughout the design process.

Chapter 4

Design Creativity Assessment Metrics in Engineering Education

The previous two chapters have brought to light the utility of creativity evaluation metrics, and in particular our TASC metric, for providing a proxy for expert ratings in creativity evaluation. However, little research exists that studies the application and implementation of these tools by engineering students on grade-dependent class projects Prior research in concept evaluation, and selection methods has largely focused on evaluating designs within the constructs of research, and generally fail to assess the abilities of these methods to be implemented by practitioners such as students, and industry designers. Although this research has provided the design community with a basis for guiding their design selection process, there remains a need for application-oriented methods and tools to utilize this knowledge. Therefore, the final study of this thesis was developed in order to test the utility of the TASC metric in an engineering design course project with 30 engineering students. The results from this study demonstrate the utility of the TASC concept-selection method in an engineering classroom, and enhance our understanding of design-creativity evaluation.

4.1 Introduction

Teresa Amabile said that “you don’t want to ghettoize creativity; you want everyone in your organization producing novel, and useful ideas” [149]. This is found to be true throughout engineering design where researchers, and practitioners are meant to continuously exchange ideas for the purpose of contributing to the forward momentum of understanding creativity in product design [103, 150]. In fact, both researchers and practitioners are now pushing for tools to quantify creativity that are rigorous, and reliable for use in both environments [151, 152]. This creates a challenge for new creativity assessment tools, and methods that may be rigorous but lack the extensive testing and proper implementation strategies to be considered reliable. Therefore, it is important to identify the strengths, and limitations of creativity metrics in order to improve our overall understanding of the design process.

One of the most efficient ways to obtain measurements of design creativity is to have designers subjectively rate, and select designs that are perceived to be the most creative. Either individuals or design teams can utilize this method in a relatively short period of time using internal heuristics [153, 154]. However, there are a number of drawbacks to this sort of assessment due to the inherent biases and lack of control for variations between designers and within teams. Personal preferences, self-monitoring effects and fixation are some of the many specific attributes that call into question the rigor and repeatability of this type of creativity assessment [155, 156]. Because of the biases inherent with these methods, several creativity metrics have been developed to mitigate designer bias through intricate models and methodologies that compare and contrast unique design features using decision-making theory [39, 138, 157]. In this way, quantitative values for a design's creativity as well as quantity and variety have been calculated using methods such those developed by Shah, Vargas-Hernandez, and Smith [24] that have become a landmark in the study of design creativity [25, 135]. Although this method exists, it is one of only a few methods currently available and it has often served as the base for other metrics perpetuating the existing limitations [25, 30]. Therefore, there is a need for developing more effective and efficient methods for practicing designers and students.

Therefore, the purpose of this final investigation was two-fold. First, we seek to understand the impact of teams evaluating self-generated design creativity using both subjective judgment and a new creativity metric as compared to more established metrics. Second, we aim to understand the impact of individual evaluations of self-generated design creativity using both subjective judgment and the creativity metrics. The results from this study are used to derive design recommendations to develop a web-based design creativity evaluation tool that can be used by both student designers and eventually industry designers. In this way, creativity can gain support throughout the design process and engineering design as a field.

4.2 Background

4.2.1 The Impact of Teams in Engineering Design Activities

Modern engineering design relies heavily on the interaction, and productivity of multiple engineers working together in teams. Researchers in engineering education frequently cite the pressure directed from industry to teach teamwork as early as possible [158-161]. For this reason, psychologists, and engineers have conducted significant studies to understand the interworking of teams while participating in creative activities such as brainstorming, and decision-making. For example, research conducted by Bechtoldt et al. [162] has shown that time pressure coupled with positive social acceptance within teams resulted in increased productivity, and originality of creative ideas generated. Knowledge about team interactions such as this can be leveraged within engineering designs tools to obtain more novel, and feasible designs. In fact, several team brainstorming tools, and techniques rely on this type of knowledge by encouraging design teams to generate as many ideas as possible without judgment [163-165].

While these findings can help to foster creative teamwork, other research has explored the complex nature of teams, and creativity that may require a pause for deeper consideration. For example, teams have been found to experience reduced brainstorming performance due to social loafing within the teams, and individual preferences against teamwork are problematic [166, 167]. Personality tests such as the Myers–Briggs Type Indicator assessment have helped to measure these individual differences, and improve problem solving, and teamwork [168, 169]. The existence of these types of effects highlight the need for caution when developing engineering design tools that could be used by teams with various compositions. Therefore, the current study seeks to understand the variability of team design creativity evaluations while using various creativity assessment tools, and methodologies.

4.2.2 The Impact of Nominal Team Interaction on Design Creativity

While the previous section outlined the benefits and limitations of interactive and diverse design teams, nominal teams that perform brainstorming individually and then meet later as a team to review the designs must also be considered. In fact there have been conflicting views found in prior research on the benefits, and limitations of group product-design brainstorming sessions, and the level of creative ideas produced [170, 171]. The factors that impact such teams have been studied in prior literature in social psychology, and engineering design [172, 173]. For example, a study conducted by Rietzschel et al. [106] looked at the performance of nominal design teams at generating, and selecting ideas. They found that, while nominal design teams were able to *generate,* and more novel design than interactive teams, both teams performed similarly at design *selection* [174]. These findings are significant in exposing how easily team composition, and organization may have an impact on certain aspects of the design process. Therefore, it is important that we study nominal design teams as they relate to creativity assessment and design selection.

Nominal design teams present an opportunity to observe shifts in preferences, and decision-making abilities from when designs are selected individually, and then as teams. For example, Sniezek and Henry [175] found that combined averages of individual judgments were less accurate than group judgments. This is important to consider when developing unified creativity assessment methods that may be used to evaluate self-generated designs where bias is already contributing to decision variability [176]. However, research by Kerr and Tindale has explored the idea of "group-based forecasting" of collective judgment from individuals, and found that although this type of "forecasting" can be used; there are factors such as the size of the group, and interaction modalities that can impact results [177]. This deeper understanding of the impact of nominal design teams in design activities contributes to the efforts of developing new engineering design tools that can effectively measure design creativity despite varying design environments.

4.2.3 Lessons Learned from Current Creativity Evaluation Methods and Tools

A key contribution of new creativity evaluation research and methods is to provide a base for application, and additional development that will increase the support of creativity throughout the design process. This is important because creativity has become increasingly integral to the success of engineering solutions, and products [178]. One of the many missions of an engineering researcher is to ultimately take findings, and provide recommendations for how they may be applied by others in their respective design efforts whether they be additional research, practice or education [152, 179]. As pivotal researchers in the study of design creativity, Shah, Vergas-Hernandez, and Smith have provided interesting questions regarding the practicality of consolidating the many measures of design creativity (e.g., quantity, quality, novelty and variety) into a single overall score [24]. In their original SVS metric, designs are decomposed, and evaluated by their ability to satisfy predefined functional requirements as a function of the number of similar ideas generated [19, 24]. In the field of engineering design, measurement granularity may be beneficial in that a single value can be effective for ease of communication but several specific values could be effective in design creativity optimization, and comparisons. For instance, computational design creativity researcher David Brown arrived at the conclusion that, despite the availability of the four measures of design creativity (i.e., SVS's measures), they remain too ambiguous for computational creativity methods [30]. Therefore, there are benefits to developing design creativity metrics that can provide both a general measure of design creativity as well as a more detailed set of design attributes.

In addition to defining the granularity of creativity evaluations, several scholars have highlighted opportunities in crowd sourcing creativity judgments, and utilizing simplified methods that reduce the workload of design raters. Research by Kurt Luther has shown that online crowdsourcing tools have an opportunity to act as catalysts for creativity, and innovation by allowing individuals of varying expertise to work together [180, 181]. As for simplifying design tools to make them accessible for both practitioners, and participants, researchers at Microsoft have developed several user-centered tools, and methods to measure product desirability and usability [37, 95]. In this way, designers are empowered to seek out designs that are not just feasible and usable, but also creative without wasting research time following ambiguous procedures, or confusing the individuals providing feedback. The results of these prior studies provide opportunities that can be addressed with new design-creativity assessment methods. Therefore, the current study seeks to understand how to design new creativity methods that can be easily implemented by design teams of varying abilities, but with comparable results.

4.3 Research Objectives

The purpose of the current study was to examine the relationship between creativity evaluations performed with both individual engineering design students, and student design teams on self-generated designs. Specifically, our study was developed to answer the following questions:

1. How do individual creativity concept evaluations compare to team evaluations? We hypothesize that there will be variability in the way individual and teams judge the creativity of design ideas because prior research has shown that interpersonal differences between teammates can be constrained during group brainstorming activities leading to more creative designs [177], while other research has shown that social loafing and halo effects may lead to reduced quality of engineering decisions [167, 182].
2. How do judgment-based evaluations compare to structured creativity metrics (TASC and SVS)? We hypothesize that since human judgment-based evaluations is based on a global understanding of creativity and structured creativity metrics are based on relative evaluations of creativity, these methods would results in a limited relationship between design evaluations [14, 25, 183]. . However, other research has supported the abilities for new creativity metrics to overcome these issues such as our TASC method presented in these proceedings [19].
3. How do individual creativity concept evaluations compare to structured creativity evaluations methods (e.g., TASC and SVS)? We hypothesize that evaluations of design creativity by the individual teammates will be influenced by individual biases and our TASC creativity assessment tool will be effective at mitigating these interpersonal and team biases. Prior research has shown that both differences in relatable experience and individual preferences can impact their judgments of design novelty and quality [59, 184].

4.4 Methodology

To answer these research questions, a study was conducted with 30 undergraduate student engineering students. This section serves to summarize the methodological approach taken in this study.

4.4.1 Participants

At the start of the study, the purpose and procedure of the study were discussed and any questions were answered. Next, implied consent (IRB) was attained. The participants in this study were undergraduate students in a first-year engineering design course at Penn State. There were 30 students (21 males, 9 females) that participated in this study from one section of the course. Within the section, three- and four-member design teams were established resulting in eight teams in total with two teams consisting of three members. Questionnaires were given out at the start of the semester that asked about student proficiencies in 3D modeling, sketching and engineering design. Teams were assigned by the instructor based these questionnaires to balance the performance of the teams

4.4.2 Procedure

At the start of the study, students participated in an in-class brainstorming session during which each team member was given approximately 20 minutes to individually generate as many ideas as they could for the following design task: "Your task 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."

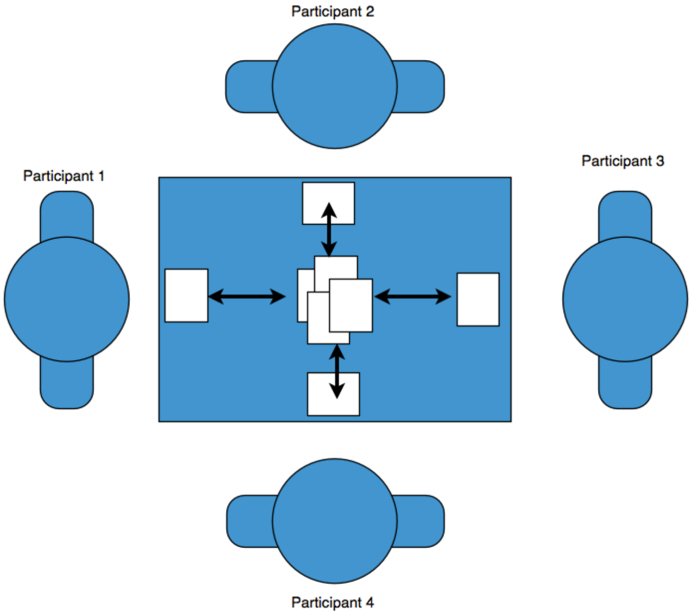


Figure 4-1 A diagram showing the process used by students to select designs to evaluate throughout the study.

During the brainstorming session, students were asked to write any notes on the sketches that would aid an outsider with understanding the concept's purpose or features. Each student was provided with sketching papers that had borders and labels to clearly distinguish the student's identification code, team name, and the specific design number. Next, the teams were asked to carefully discuss the ideas developed by the team, and sort the design sketches into two piles: "consider", any design they wanted to consider for further development, and "not consider," designs the team no longer wanted to consider as part of the process. After this, teams were instructed to go through the designs in the "consider" pile, and rank them from most likely to develop to least likely to develop as a design for their final class project Post-it notes with the ranking for each design were placed on corresponding sketch, and digital pictures of the final rankings were taken by the experimenters. Participants were not able to see these ratings again until the end of the study period. Overall, this process of team consensus rankings of the design took 30 minutes.

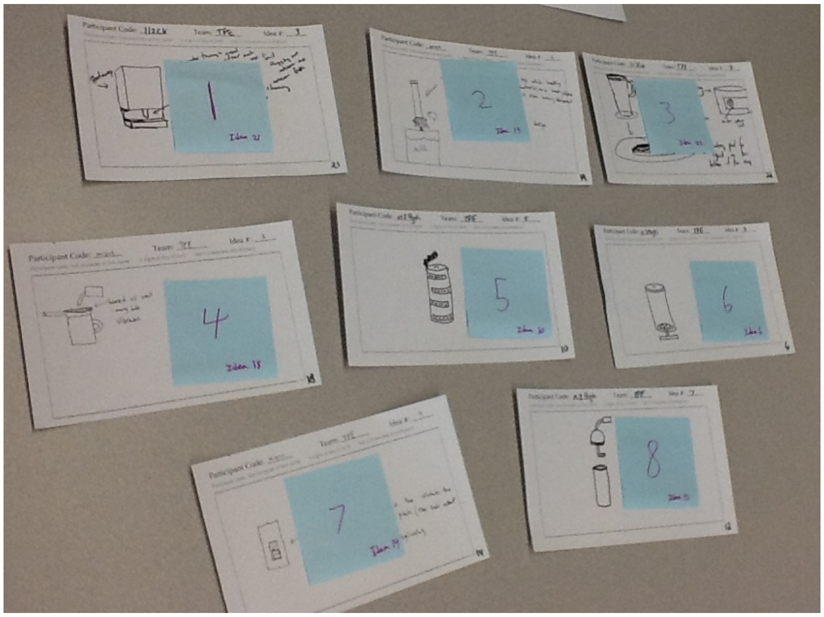
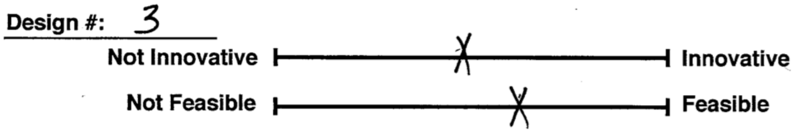


Figure 4-2 An example of teams ranking designs using Post-it notes with comments

Following the team design evaluation activity, Design Assessment Toolkit (DAT) packets were distributed each student. The DAT is a toolkit composed of two parts developed to help students measure design creativity based on their individual judgment, and our TASC method based on word selections, and semantic similarity. A copy of the DAT can be found in the Appendix. Next, each student was instructed to randomly select one of their team's sketches at the center of their table, and to complete the individual judgment portion of the DAT, see

Figure 4-1 and Figure 4-3 for examples. In this way, each member of the team evaluated every design produced by the team. The students were given approximately 20 minutes to complete this task.


Design #: 3

Not Innovative — Innovative

Not Feasible — Feasible

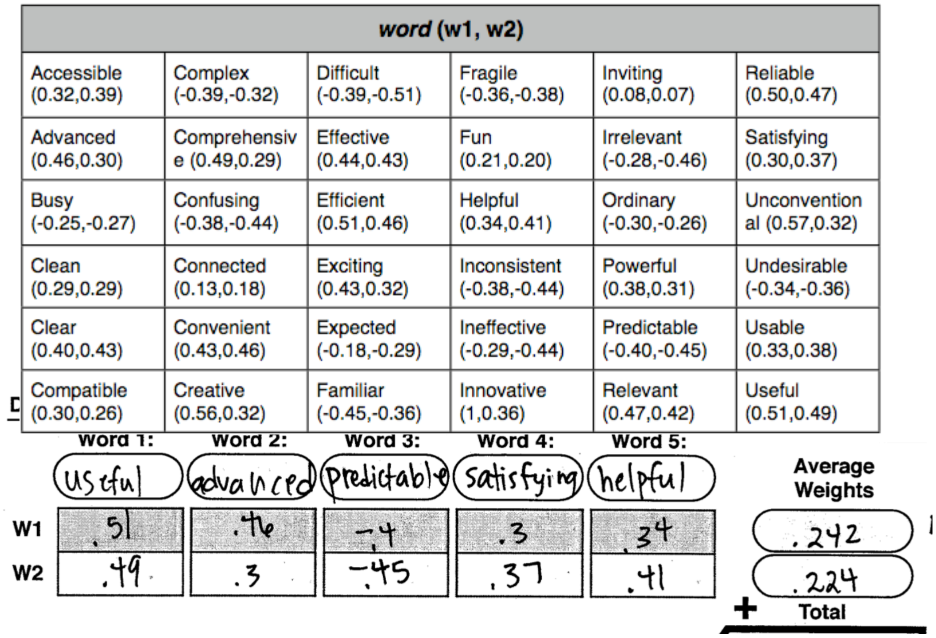
Figure 4-3 Individual judgment of design creativity using continuous scales.

Once the designs were evaluated using individual judgment, the students were instructed to move on to the second part of their DAT packets for which they were provided an index card with a table of 36 words. Once again, each teammate was instructed to randomly select one of the team's designs, and choose three to five words from the index cards, and fill them into the DAT Exercise Form. This was repeated by every teammate until they evaluated all of the designs generated by the team. Upon completion, students were provided with new index cards with both the original words, and two sets of numeric weights for each word, as seen in Figure 4-4. With the word list and weights in hand, the students were instructed to match the words they selected on the DAT Exercise Form with their corresponding weights and fill in the blanks for all designs, see Figure 4-3 for an example. Once this was completed, students were instructed to compute the average word weights, and total scores for each design. The students also completed a short questionnaire regarding the evaluation methods (see Appendix for questions). Afterward, the DAT Exercise Forms were collected from each team, and the study concluded.

From these forms, Team TASC evaluations were tabulated, and returned to each team for their own records. An example of a completed Team TASC evaluation score sheet is presented in Figure 4-5. All DAT packets were scanned and returned to students. Although students were provided with both the individual judgments, and TASC score summaries, they were not required to use the information moving forward in their projects. In total, 60 designs were evaluated using each method, and all TASC calculations were checked for arithmetic errors before data analysis.

4.4.3 Creativity Assessment Tools and Metrics

In order to investigate the differences between individual and group evaluations, several creativity valuation metrics were developed and utilized in this study. In the following sections, each method will be described and relevant calculations will be described in detail.



word (w1, w2)					
Accessible (0.32,0.39)	Complex (-0.39,-0.32)	Difficult (-0.39,-0.51)	Fragile (-0.36,-0.38)	Inviting (0.08,0.07)	Reliable (0.50,0.47)
Advanced (0.46,0.30)	Comprehensive (0.49,0.29)	Effective (0.44,0.43)	Fun (0.21,0.20)	Irrelevant (-0.28,-0.46)	Satisfying (0.30,0.37)
Busy (-0.25,-0.27)	Confusing (-0.38,-0.44)	Efficient (0.51,0.46)	Helpful (0.34,0.41)	Ordinary (-0.30,-0.26)	Unconventional (0.57,0.32)
Clean (0.29,0.29)	Connected (0.13,0.18)	Exciting (0.43,0.32)	Inconsistent (-0.38,-0.44)	Powerful (0.38,0.31)	Undesirable (-0.34,-0.36)
Clear (0.40,0.43)	Convenient (0.43,0.46)	Expected (-0.18,-0.29)	Ineffective (-0.29,-0.44)	Predictable (-0.40,-0.45)	Usable (0.33,0.38)
Compatible (0.30,0.26)	Creative (0.56,0.32)	Familiar (-0.45,-0.36)	Innovative (1,0.36)	Relevant (0.47,0.42)	Useful (0.51,0.49)

	Word 1:	Word 2:	Word 3:	Word 4:	Word 5:	Average Weights
	useful	advanced	predictable	satisfying	helpful	
W1	.51	.46	-.4	.3	.34	.242
W2	.49	.3	-.45	.37	.41	.224
						+ Total

Figure 4-4 Individual TASC evaluation of design creativity

SVS Design Creativity

To quantify the overall design creativity, the SVE method to calculate design novelty and quality was utilized [24]. The specific measures of design novelty, and quality were used to calculate an overall SVS creativity score because prior research has defined creative designs as fitting both criteria of being novel, and feasible [17, 185]. In this way, design novelty is calculated by looking at the relative uniqueness of a design's separate features within a given set of designs. For example, designs with more common features (i.e., shape, button placement, attachments) are considered to have less novelty. Design quality on the other hand is defined as the how well an idea is able to meet the design specifications required [24]. This was evaluated by answering a set of questions that asked if a design is technically feasible, easy to execute and completes the desired task. Both of these measures result in scores between 0, and 1 with 1 being considered the most novel or most feasible (of quality). (For more in-depth discussion of design novel, and quality calculations see [186] and equations below.) By adding these two values together, an overall SVS creativity score is obtained.

Feature Novelty f_i

The method of computing f_i is

$$f_i = \frac{T - C_i}{T}, \qquad (4\text{-}1)$$

where T is the total number of concepts generated for each category of design feature, and C is the total number designs rated for each feature.

Design Novelty D_j

The method of computing D_j is

$$D_j = \frac{\sum f_k}{\sum f_i}, \qquad (4\text{-}2)$$

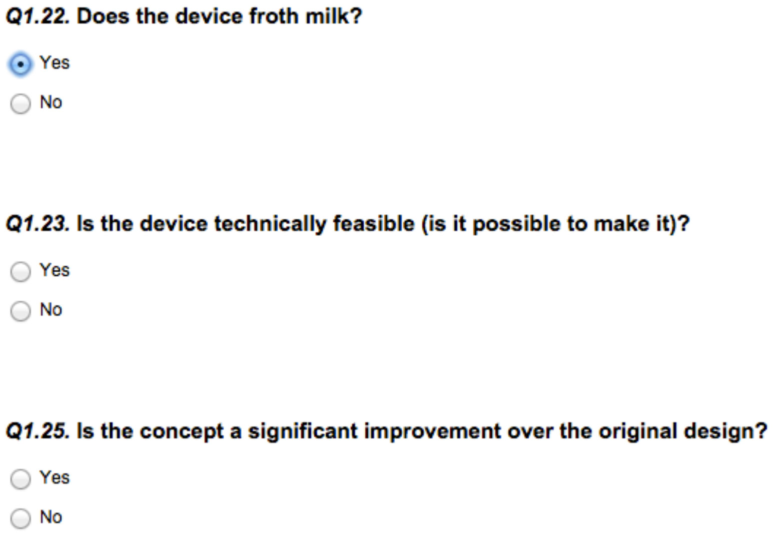
where f_k is the feature novelty of a feature from the original design and f_i is the total novelty of all features that were assessed.

Design Quality Q_i

Design quality is calculated based on rater responses to three questions about the design's ability to perform the intended task. An example of the questions asked is provided in Figure 4-5 and design quality is calculated from the responses to the questions.

The method of computing design quality, Q_i is

$$Q_i = \frac{\sum_{k=1}^{3} q_k}{3}, \qquad (4\text{-}3)$$



Q1.22. Does the device froth milk?

● Yes

○ No

Q1.23. Is the device technically feasible (is it possible to make it)?

○ Yes

○ No

Q1.25. Is the concept a significant improvement over the original design?

○ Yes

○ No

Figure 4-5 Example quality questions using the SVS method.

where q_k is the answer to the kth quality question. If "yes" is answered then q_k is equal to 1, and when the answer is "no" then it is equal to 0 [187].

SVS Creativity C_j

SVS creativity is the combination of the novelty, and quality metrics that were originally developed by Shah, Vargas-Hernandez, and Smith [24]. The method of computing C_j is

$$C_j = D_j + Q_j, \qquad (4\text{-}4)$$

where D_j is design novelty for idea j, and Q_j is design quality for idea j.

Individual Creativity Concept Evaluations

In this study, individual creativity concept evaluations were calculated by measuring the markings on two scales regarding design novelty, and feasibility, see Figure 4-2. For designs with little novelty, markings range between the extremes of "not innovative" and "innovative". As for design feasibility, markings are measured likewise between the extremes of "not feasible" and "feasible". With the respective measurements from the lower extreme, a ratio can be calculated by dividing the marked length by the total length of the scale. This was completed for both scales. Finally, to obtain a creativity score for a design, the novelty, and quality ratios were added together.

Team Creativity Concept Evaluations

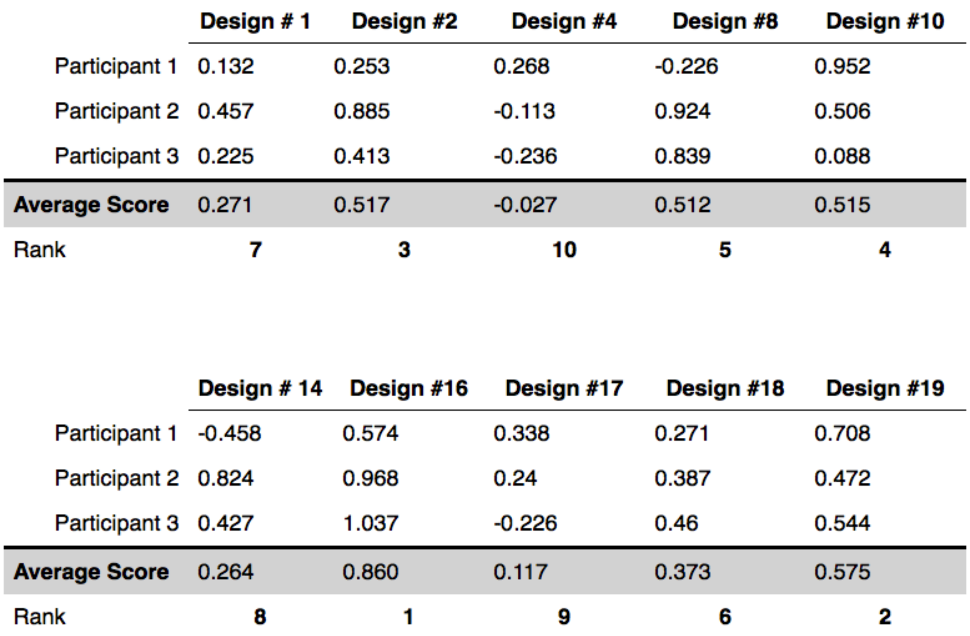
In this study, the team creativity concept evaluations are calculated by aggregating the individual creativity evaluations from each teammate. For each design, the individual creativity evaluations of the teammates are averaged together to obtain a single team concept creativity evaluation.

Informal Team Discussion Creativity Concept Evaluations

Informal team judgments of design creativity are evaluations of design creativity obtained from teammates conversing in a group, and ranking designs from most appropriate to least appropriate using Post-it notes. An example of team judgments of design creativity is shown in Figure 4-1.

TASC Method for Evaluating Design Creativity

The TASC method was developed over a serious of prior studies to use word selection from raters, and semantic similarity (word relatedness) to calculate design creativity. This involves having raters choose three to five words from a set list for each design, and matching them to two semantic weights, see Figure 4-3. The sematic weights are a measure of the relatedness between each of the words in the list and the words "innovative" and "feasible" obtained from using an online toolkit called DISCO [99]. By averaging the "innovative" semantic weights for each word, and doing the same for "feasible" semantic weights, we can obtain novelty and quality scores. Finally, the overall creativity score for each design is calculated by adding both scores together. In teams, the TASC creativity scores by each teammate can be averaged to obtain singular scores for each design, see Figure 4-6 for an example of team TASC evaluations.



	Design # 1	Design #2	Design #4	Design #8	Design #10
Participant 1	0.132	0.253	0.268	-0.226	0.952
Participant 2	0.457	0.885	-0.113	0.924	0.506
Participant 3	0.225	0.413	-0.236	0.839	0.088
Average Score	0.271	0.517	-0.027	0.512	0.515
Rank	**7**	**3**	**10**	**5**	**4**

	Design # 14	Design #16	Design #17	Design #18	Design #19
Participant 1	-0.458	0.574	0.338	0.271	0.708
Participant 2	0.824	0.968	0.24	0.387	0.472
Participant 3	0.427	1.037	-0.226	0.46	0.544
Average Score	0.264	0.860	0.117	0.373	0.575
Rank	**8**	**1**	**9**	**6**	**2**

Figure 4-6 Team TASC evaluation of design creativity

4.4.4 Statistical Analysis

In order to test our hypotheses, creativity scores were calculated from individual judgments, team judgments, our TASC method and *post hoc* using the SVS method with two independent raters. In order to compare these different methods of evaluating design creativity, creativity scores were normalized between 0 and 1 with 1 being the most creative design. These scores were then ranked for analysis using partial correlations first controlling for team number and then for participant number. SPSS v.22 was used to analyze the data with a significance level of 0.05.

4.5 Results and Discussion

The purpose of this research was to investigate the relationship between creativity evaluations performed with individual engineering design students, and student design teams on self-generated designs. Therefore, these interactions were analyzed in three phases. The first phase was to explore the interaction between team creativity concept evaluations, and informal creativity concept evaluations. The second phase was performed in order to understand the relationship between these the evaluations in Phase 2, and the TASC, and SVS creativity evaluations. Finally, a third analysis was performed in order to understand the relationships between individual creativity concept evaluations, informal team creativity concept evaluations, individual TASC evaluations, and SVS evaluations.

The TASC method results allow for a review of commonly selected words and summary statistics regarding the number of words ultimately selected by the engineering design students. On average, students selected 3.97 words for each design (SD = 0.87) with a range from 3 and 5 words as instructed throughout the study. However, it was found that out of 30 students only three were found to select the minimum of 3 words for each design. This finding provides support for the significant involvement and consideration taken by the students during the study. The most frequently chosen words in the TASC method include “innovative”, “creative”, “useful”, “effective” and “usable.” However, the words chosen the least include “connected”, “inconsistent”, “inviting”, “comprehensive” and “irrelevant.” These results provide a basis for understanding the contextual use of the adjectives by students to describe their designs and their perception of design features.

4.5.1 Relationship between Team Creativity Concept Evaluations and Informal Team Evaluations

To examine the relationship between creativity evaluations using team creativity concept evaluations, and informal team creativity concept evaluations, partial correlation analyses were performed controlling for the design teams performing the evaluations. We hypothesized that these two methods of evaluating design creativity would differ due to the individual differences in teammate preferences as compared to evaluations performed together through discussion. This relationship was then subjected to a first-order partial correlation in order to explore the relationship controlling for the effects of the design team. The first order correlation was found to be statistically significant, $r = 0.59$, $p < 0.001$.

The results indicate that the team creativity concept evaluations were significantly related to the informal team creativity concept evaluations. In particular, the relationship between these evaluations was of moderate correlation strength but of extremely high statistical significance. This suggests that, by taking an average of individual judgments by nominal teams, we are able to obtain a moderate understanding of how the individuals will evaluate design while they are in person together evaluating design creativity. This finding may be due to the fact that the team creativity concept evaluations were performed after the informal team creativity concept evaluations. Therefore, these findings may represent some of the informal team creativity concept evaluations influencing the team creativity concept evaluations despite the fact that teams were not able to directly compare evaluation results.

4.5.2 The Abilities of Creativity Evaluation Tools Minimizing the Impact of Interpersonal Judgments

Our second research question was developed to investigate our team TASC evaluations and the SVS method. This relationship was then subjected to a first-order partial correlation in order to explore the relationship controlling for the effects of the design teams. The first-order correlations found significant relationships between all relationships with the exception of the SVS method, see Figure 4-7 and 4-8 for a summary of results. The relationship between team creativity concept evaluations and our team TASC evaluations were found to have the following first-order partial correlation, $r = 0.58$, $p < 0.001$. The relationship results also showed a relationship between team judgments and our team TASC evaluations with the following first-order partial correlation of $r = 0.51$, $p < 0.001$ while controlling for the design team.

These results suggest that our team TASC evaluations of design creativity are able to capture similar constructs of creativity throughout the design teams and how they perceive design creativity. However, the widely adopted SVS method failed to exhibit any significant partial correlations with all the other evaluation methods used in this study including our TASC evaluations. These results indicate that, while our team TASC evaluations are tapping into similar constructs of perceived creativity by teams, the SVS method does not appear to be doing so. Therefore, the relationships between these methods should be explored on the individual evaluator level to understand their granularity.

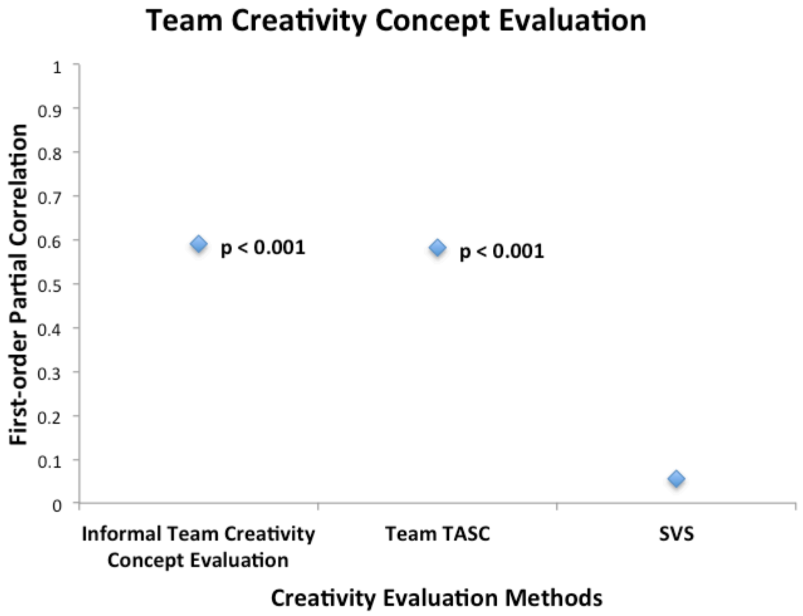


Figure 4-7 Summary of first-order partial correlation relationships between team creativity concept evaluations and other creativity evaluation methods

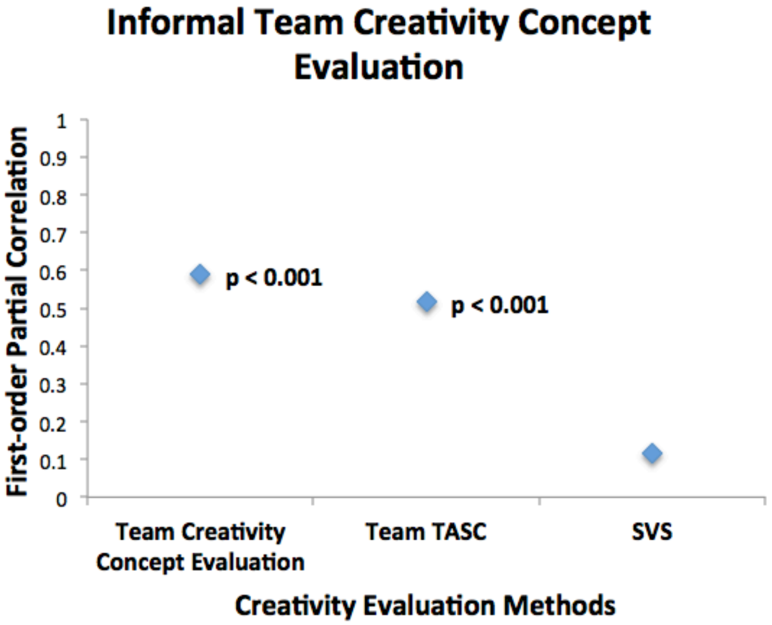


Figure 4-8 Summary of first-order partial correlation relationships between informal team creativity concept evaluations and other creativity evaluation methods

4.5.3 Individual Creativity Evaluations

In order to investigate if relationships exist between individual creativity concept evaluations, informal team creativity concept evaluations, individual TASC evaluations and the SVS method, a final series of first-order partial correlations were computed while controlling for the individual student. The results revealed that the relationship between our individual TASC evaluations and the SVS evaluations were the only pair lacking significance, see Figure 4-9 and 4-10 for a summary of results. The relationship between individual creativity concept evaluations and our individual TASC evaluations were found to be significant at the $r = 0.46$, $p < 0.001$ level while controlling for the effects of individual students. This suggests that our TASC method is tapping into similar constructs of creativity as the students. The relationship between individual creativity concept evaluations and the SVS evaluations was also found to be significant, but to a lesser extent at $r = 0.13$, $p < 0.05$. These results suggest that at the individual evaluation level, our TASC method maintains a moderate relationship with the way designers are evaluating design creativity. Conversely, these results also highlight the limitations of the widely adopted SVS method and its use for evaluating design creativity as a design team would. Therefore, the results direct us to new knowledge regarding the relationships between human judgment and creativity evaluation methods and design creativity within teams and between individuals.

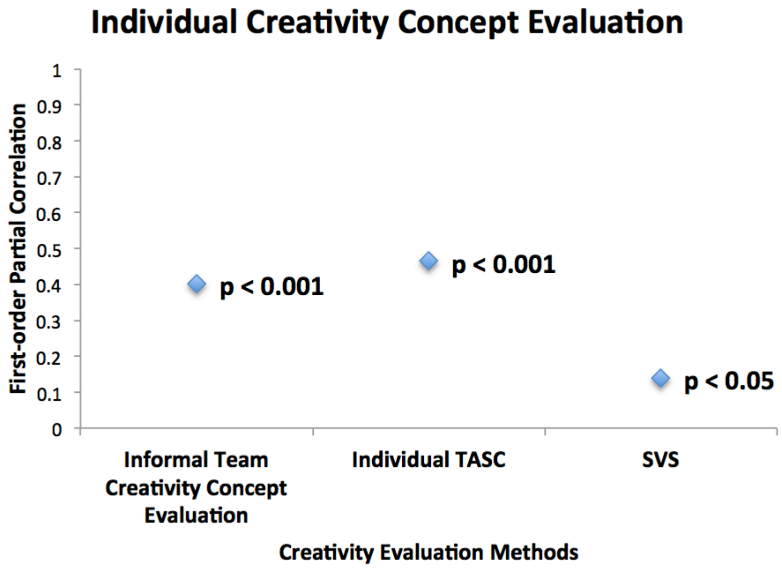


Figure 4-9 Summary of first-order partial correlation relationships between individual creativity concept evaluations and other creativity evaluation methods

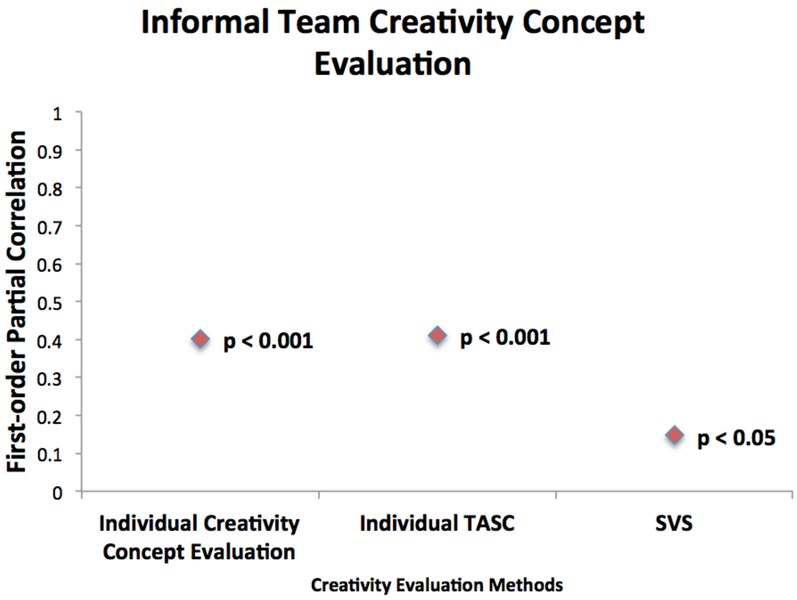


Figure 4-10 Summary of first-order partial correlation relationships between informal team creativity concept evaluations and individual creativity evaluation methods

4.5.4 Summary of Word Selections and Student Feedback

The findings from the design evaluations are extended through the questionnaire feedback from students after completing the entire study. When asked about preferred method of evaluating their designs, 40% (12 students) preferred the line marking method, 43% (13 students) preferred our TASC method and 17% (5 students) either preferred using either both methods or neither method. However, when asked if our TASC method altered their thought process during evaluations, 57% (17 students) agreed that it did and 33% (10 students) stated it did not. Three students did not provide a clear response as to which method they preferred. These results show that although students were split relatively evenly between their creativity evaluation method preferences, our TASC method was impactful, and enabled students to consider designs from a broader perspective. In addition, ten of the students that preferred the method of using the individual judgment method with line markings also reported that the TASC method altered their thought processes. For example, student 24 stated that "I kept viewing it from an engineers perspective, and the words made me view it from a customer perspective."

This feedback provides insights for the development of interactive concept evaluation, and selection methods that encourage designers to consider designs more thoughtfully.

4.6 Discussion

These findings lead us back to the purpose of this study, which was to explore the relationship between individual, and group judgments of self-generated designs using creativity evaluation methods. Our main results were as follows:

- There is a relationship between team creativity concept evaluations, and informal team creativity concept evaluations, but this relationship may be due to the fact that the designs were previously evaluated together in teams leading to some priming effects;
- Our TASC creativity evaluation method is able to tap into similar constructs of design creativity as both individual, and team creativity evaluations, but the correlation coefficient is rather low (< 0.6) demonstrating variation in these methods and the potential for mitigating human biases; and
- The SVS method that has been widely adopted in design creativity research may not be capturing creativity, as it is perceived by design teams.

Prior to the study, it was hypothesized that there would be differences between informal team creativity concept evaluations, and informal team creativity concept evaluations because prior studies have shown that in brainstorming tasks, teams were not necessarily generating more creative designs than nominal teams as others have suggested [106]. However, the current study was conducted to understand how individual differences in creativity evaluations could be influenced by informal team evaluations of creativity. The results from the study indicate that, despite prior research's findings regarding the limitations of forecasting team creativity from individuals, the team creativity concept evaluations using averaged individual creativity concept evaluations do have a relationship with informal team judgments [177]. Although, as described earlier, the informal team creativity concept evaluations were performed prior to the team creativity concept evaluations. This ordering of the evaluations could have resulted in some priming effects. However, these results maintain a margin for variability included within in the moderate relationship found between their evaluations. So, although these two methods have some variability in their evaluations, they appear to be tapping into similar constructs of creativity.

Building on this finding, the results revealed that our TASC evaluation method shares a relationship with both team creativity concept evaluations, and informal team creativity concept evaluations. This shows promise for the use, and application of the TASC to measure design creativity with individual designers and development teams. This is supported by prior research at Microsoft with the Desirability Toolkit, and the subsequent Product Reaction Cards that have cited successes of using word selections to obtain user feedback regarding intangible measures such as desirability or in our case creativity by practitioners [37]. Therefore, the results have shown a link between how we perceive creativity, and the way our TASC method is able to obtain quantitative evaluations.

Conversely, the results also revealed some of the limitations of creativity metrics often cited in engineering design research in academia. With regards to team judgment of creativity, the results did not find any significant overlap between the team judgments of design creativity and the SVS evaluations. A criticism of the SVS method has been that it lacks appropriate evaluation granularity due, and although a feature may be unique to the design set it may not be a creative design overall [25, 30]. Another problem facing this method is implementation within a classroom or fast paced development in industry. To calculate the SVS scores for this study, a customized milk frother benchmarking handbook had to be developed, and required hours to complete effectively. Finally, when the handbook was completed each design needed to be evaluated following the handbook. Ultimately, the amount of time and effort required to obtain the SVS evaluations were disproportionate to their ability to evaluate design creativity in the context of a classroom.

Although the SVS, method may be removing biases, and subjectivity through its meticulous feature-level evaluations and functional questionnaires, it appears to be conflicting with human perception of what something means to be creative. However, the results from the study indicate that our TASC method was able to display significant relationships with both measures of human’s perceived judgment of creativity. It appears to be tapping into similar paradigms of design creativity that could find middle ground between the wisdom of the crowd and individual-based knowledge. In summary, our study suggests that there is a need for design-creativity evaluation methods that can help to mediate the benefits of human judgment with the rigors of structured methods such as the SVS method, especially within a classroom.

4.7 Conclusions

Overall, the results of this study show that there is a need for creativity evaluation tools that are both easy to implement, and compliant with the definition of design creativity (i.e., novel and feasible). This has important implications for engineering design research, education, and practice, where there is an increasing demand for products that are not just technologically sound but, even more so, creative in their incarnation [188, 189].

The results from this study have important implications for engineering education, and practice in industry. The results show that our TASC method may be a means to reconcile differences between individual judgments, and group judgments of design creativity, while being built on a structured framework that enables accessibility, and repeatability. However, there exist several limitations that are important to note. First, our study was performed with one class of undergraduate engineering students evaluating their class project designs. It would be beneficial to see if these results are similar with other undergraduate engineering classes. In addition, the design problem was the same for each of the design teams, and future studies can explore the impact of team specific design problems, and their impact creativity evaluations. Additional future work will focus on fine-tuning the TASC method into a printable toolkit for classroom use, and an online tool for anyone with an Internet connection to use. In this way, students, and designers alike will be able to consider design creativity quickly, and easily throughout the design process. This contributes to the on-going goal of improving engineering design education and developing tools to enable creative engineering design.

Chapter 5

Summary, Contributions, and Future Work

Design creativity evaluation is an essential part of engineering design education, and practice. The research presented in this thesis was developed to explore the utility, opportunities, and implications of design-creativity evaluation methods used in academia, industry and engineering education. The first manuscript of this thesis, presented in Chapter 2, investigates the utility of word selections, and semantic similarity as a method for evaluating design creativity. The second manuscript of this thesis, found in Chapter 3, explores the impact of experience, and biases on current methods of concept evaluation including our word selection method. The third, and final manuscript, found in Chapter 4, examines the impact of student students using creativity evaluation methods on self-generated designs. A summary of these studies is described in Chapter 1.

The findings in this thesis provide quantitative support for the opportunity for creativity evaluation tools to encourage creativity throughout the design processes, and to expand our knowledge on creativity in engineering design. In addition, the development of a new creativity evaluation tool is investigated for its use by students, and practicing engineering designers. This research contributes to our understanding of design creativity, and concept selection in early phases of product design and the follow contributions to work in this area:

1. This research adds to our understanding of using word selection, and semantic similarity to evaluate design creativity. The results of the studies conducted in this thesis indicate that word selection enable design-space-independent evaluations of design creativity. These results indicate that it may be possible to compare, and contrast designs from dissimilar design problems, and its advancement over prior evaluations of design creativity.
2. This research provides empirical evidence that supports the use of novice design engineers as proxies for expert experience as raters for design creativity. The studies conducted in this thesis demonstrate that despite biases, and levels of experience, individuals, and groups are tapping into similar constructs of design creativity during evaluations.

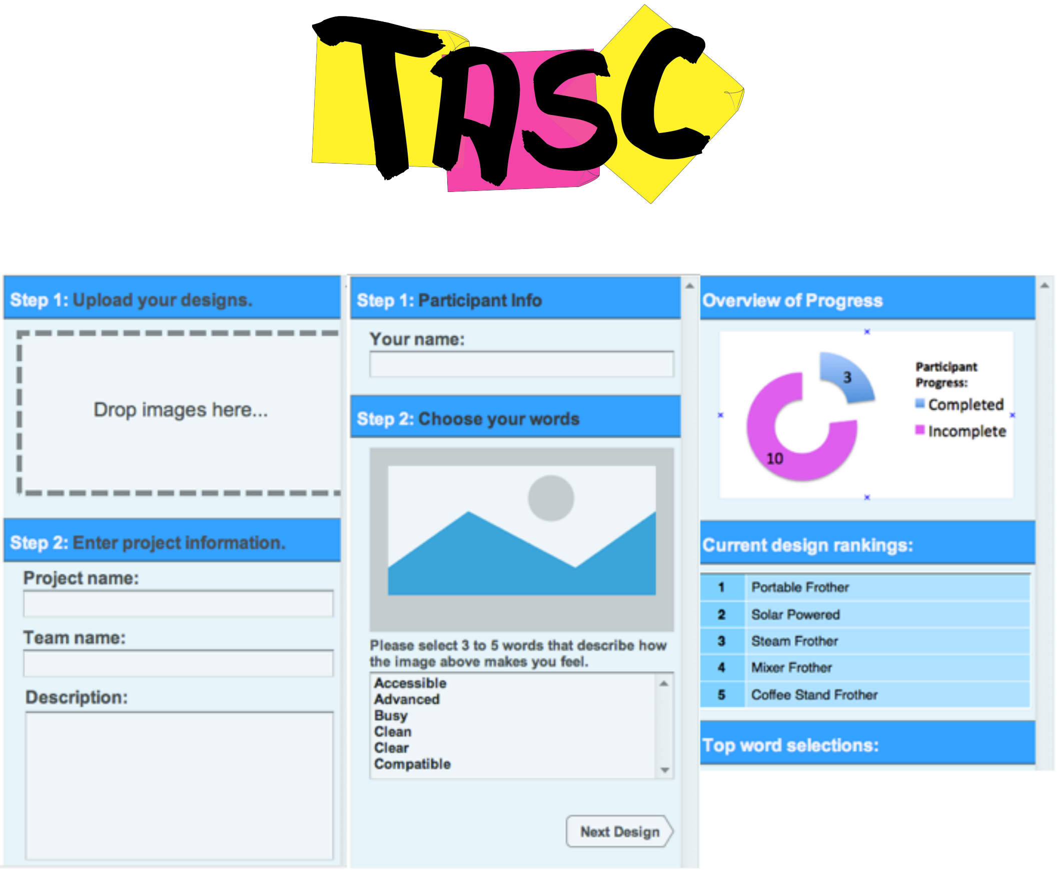
3. The results from this research provide insights into the benefits and limitations of current creativity evaluation metrics. Although prior methods attempt to reduce subjectivity by evaluating designs based on unique features, they fail to appropriately account for design novelty and quality as equal parts of a design's overall creativity.
4. This research provide further insights into the development of future creativity-evaluation methods that can be used effectively by individuals as well as groups despite their various individual differences.

This research also provides important evidence to assist with the development of a printable toolkit, and web-based design creativity evaluation tool based on our TASC method described in Chapter 4 of this thesis. While this research focused on the opportunities, and limitations of creativity evaluations tools, the results can contribute to the development of other new design-creativity tools for engineering education, and engineering practice. Thus, future research is needed to enable the usability, and accessibility of these design-creativity tools.

5.1 On-going and Future Work

In an effort to design accessible, and usable creativity evaluation tools, our research has resulted in the development of a web-based TASC concept creativity evaluation, and selection tool. The goal of this tool is to encourage both students, and professionals to truly consider creative designs throughout the development process. The future TASC web-app will allow any individual to start a design-selection project at a moment's notice by uploading pictures of their team's design sketches as the project's administrator, see Figure 5-1 for screenshots.

After uploading the pictures, the project's administrator will be presented with a page to insert the email addresses of each of their teammates who will be participating in the design evaluation. When the project administrator clicks submit, the each member of a team will receive an email with their access link to begin evaluating the designs. The designs will then show one page at a time with a list of words below the image. A participant will be prompted to select three to five words from the supplied list to describe the design that is currently being presented on the screen. Once the words are selected, they can move on to the next design until all designs have been evaluated. During this time, the project's administrators can follow the administrator link to a site displaying real-time status updates. The administrator can track the progress of evaluations as well as the current rankings of the designs. Once the team evaluates all of the designs, they can print out a full report that contains the ranks of the designs, the words used to select the designs and the creativity, novelty, and feasibility scores for each design. This web application require no training at all and can be use to evaluate designs at various stages of the development process. This creativity evaluation and selection tool will be made available to the public in the Spring of 2015.



Step 1: Upload your designs.

Drop images here...

Step 2: Enter project information.

Project name:

Team name:

Description:

Step 1: Participant Info

Your name:

Step 2: Choose your words

Please select 3 to 5 words that describe how the image above makes you feel.

Accessible
Advanced
Busy
Clean
Clear
Compatible

Overview of Progress

Current design rankings:

1	Portable Frother
2	Solar Powered
3	Steam Frother
4	Mixer Frother
5	Coffee Stand Frother

Top word selections:

Figure 5-1 Wireframe designs of mobile TASC web-app. Start in top right and go clockwise: Upload, register team information, evaluate designs, overview of results

5.2 Limitations & Opportunities

While this thesis has explored the utility of using word selections, and semantic similarity to improve the efficiency, and accessibility of concept-creativity evaluations, there exist limitations that are important to note. The meaning of words and their interpretation by engineering designer can be influenced by the context of the design problem. In this thesis' studies, the variability of contextual meaning on words was not explored. It is possible that additional exploration of word context and interpretation within the design space could impact creativity evaluations. In addition to word context, the semantic weight evaluations were also limited due to the available technology. Future work should explore the use of other semantic-similarity evaluation software that may become available for public use and could provide semantic weights that could increase the accuracy of our resulting TASC evaluations with respect to human perception of creativity. As the perceptual meanings of adjectives change due to culture and common use, the adjectives used to describe concept sketches may also require modifications and addition over the passage of time. While Chapter 4 of this thesis explored the use of our TASC method within the context of a classroom, there remains an opportunity to explore its use in industry implementation with more complex design concepts. This also suggests an opportunity to explore the use of concept creativity evaluations by interdisciplinary design teams with varying levels of contextual knowledge (i.e., marketing, business, and engineering). In addition, the TASC web-application will require usability testing to ensure that it provides a holistic experience for both practicing and student design engineers.

5.3 Conclusions

This thesis has resulted in the development of the TASC web-application to enable the support of design creativity throughout the development process. As for the studies described in this thesis, they have resulted in a number of publications that will contribute to our understanding of creativity overall and the method used to evaluate design creativity. This thesis and its components have shown there is a need for more accurate and accessible concept-creativity tools and methods to support the creative process in practice. Ultimately, the findings from this thesis will be made available to design engineers in academia and industry and provide support for the development of creative and innovative products.

References

[1] 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.
[2] Paulus, P. B., 2000, "Groups, Teams, and Creativity: The Creative Potential of Idea-Generating Groups," Applied Psychology: An International Review, 49(2), pp. 237-262.
[3] Nonaka, I., and Peltokorpi, V., 2006, "Knowledge-based view of radical innovation: Toyota Prius case," Innovation, Science, and Institutional Change: A Research Handbook, pp. 88-104.
[4] Crozet, Y., 2010, "Driving forces of innovation in the transport sector."
[5] Saaty, T. L., 2008, "Decision making with the analytic hierarchy process," International Journal of Services Sciences(1), pp. 83-98.
[6] Frey, D., Herder, P., Wijnia, Y., Subrahmanian, E., Katsikopoulos, K., and Clausing, D., 2009, "The Pugh controlled convergence method: model-based evaluation and implications for design theory," Research in Engineering Design, 20(1), pp. 41-58.
[7] Pugh, S., 1991, Titak design- integrated methods for successful product engineering, Addison-Wesley Publishers Ltd., Strathclyde.
[8] Akao, Y., 1994, "Development History of Quality Function Deployment," The Customer Driven Approach to Quality Planning and Deployment, Asian Productivity Organization, Minato, Tokyo, p. 339.
[9] 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.
[10] Ford, C. M., and Gioia, D. A., 2000, "Factors Influencing Creativity in the Domain of Managerial Decision Making," Journal of Management, 26(4), pp. 705-732.
[11] 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.
[12] O'Quin, K., and Besemer, S. P., 2006, "Using the Creative Product Semantic Scale as a Metric for Results ‐ Oriented Business," Creativity and Innovation Management, 15(1), pp. 34-44.
[13] 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.
[14] Sarkar, P., and Chakrabarti, A., 2011, "Assessing design creativity," Design Studies, 32(4), pp. 348-383.
[15] Verganti, R., 2013, Design driven innovation: changing the rules of competition by radically innovating what things mean, Harvard Business Press.
[16] Simonton, D. K., 2013, "What is a creative idea? Little-c versus Big-C creativity," Handbook of research on creativity, pp. 69-83.
[17] Runco, M. A., and Jaeger, G. J., 2012, "The standard definition of creativity," Creativity Research Journal, 24(1), pp. 92-96.
[18] Sternberg, R. J., and Lubart, T. I., 1999, "The concept of creativity: Prospects and paradigms," Handbook of creativity, 1, pp. 3-15.
[19] 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.

[20] Okugan, G. E., and Tauhid, S., 2008, "Concept Selection Methods – A Literature Review from 1980 to 2008," International Journal of Design Engineering, 1(3), pp. 243-277.
[21] 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.
[22] Tornatzky, L. G., and Klein, K. J., 1982, "Innovation characteristics and innovation adoption-implementation: A meta-analysis of findings," IEEE Transactions on engineering management, 29(1), pp. 28-45.
[23] Simon, H. A., 1990, "Invariants of human behavior," Annual review of psychology, 41(1), pp. 1-20.
[24] Shah, J. J., Vargas-Hernandez, N., and Smith, S. M., 2003, "Metrics for Measuring Ideation Effectiveness," Design Studies, 24(1), pp. 111-134.
[25] Nelson BA, Y. J., 2009, "Refined metrics for measuring ideation effectiveness," Design Studies, 30, pp. 737-743.
[26] Chandrasekaran, B., 1994, "Functional representation and causal processes," Advances in computers, 38(73-143).
[27] Pugh, S., 1996, Creating Innovative Products Using Total Design, Addison-Wesley Longman Publishing Co., Inc., Boston, MA.
[28] López-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.
[29] Sarkar, P., and Chakrabarti, A., 2008, "The effect of representation of triggers on design outcomes," Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 22(02), pp. 101-116.
[30] 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).
[31] Reich, Y., 2010, "My method is better!," Research in engineering design, 21(3), pp. 137-142.
[32] Sun, G., Yao, S., and Carretero, J. A., 2014, "Comparing Cognitive Efficiency of Experienced and Inexperienced Designers in Conceptual Design Processes," Journal of Cognitive Engineering and Decision Making, p. 1555343414540172.
[33] Ericsson, K. A., 2006, "The influence of experience and deliberate practice on the development of superior expert performance," The Cambridge handbook of expertise and expert performance, pp. 683-703.
[34] 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.
[35] Brown, D., 2013, "Developing computational design creativity systems," International Journal of Design Creativity and Innovation, 1(1), pp. 43-55.
[36] 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.
[37] 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.
[38] 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.
[39] Okudan, G. E., and Tauhid, S., 2008, "Concept selection methods–a literature review from 1980 to 2008," International Journal of Design Engineering, 1(3), pp. 243-277.

[40] 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.
[41] Baddeley, A. D., 2002, "Is working memory still working?," European psychologist, 7(2), p. 85.
[42] McGourty, J., Dominick, P., and Reilly, R. R., "Incorporating student peer review and feedback into the assessment process," Proc. 2013 IEEE Frontiers in Education Conference (FIE), IEEE, pp. 14-18.
[43] Hayes, J. H., Lethbridge, T. C., and Port, D., "Evaluating individual contribution toward group software engineering projects," Proc. Proceedings of the 25th International Conference on Software Engineering, IEEE Computer Society, pp. 622-627.
[44] Dahl, D. W., and Hoeffler, S., 2004, "Visualizing the self: Exploring the potential benefits and drawbacks for new product evaluation," Journal of Product Innovation Management, 21(4), pp. 259-267.
[45] 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 Conference 26th Annual Conference on Design Theory and Methodology, ASME, Buffalo, NY.
[46] Osborn, A. F., 1963, Applied Imagination: Principles and procedures of creative thinking, Scribeners and Sons, New York.
[47] Yang, M. C., 2009, "Observations on concept generation and sketching in engineering design," Research in Engineering Design, 20(1), pp. 1-11.
[48] 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.
[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] Sharmin, M., Coats, C., and Bailey, B. P., "Understanding Knowledge Management Practices for Early Design Activity and Its Implications for Reuse," Proc. Proc. of the SIGCHI 2009.
[51] Burrows, P., 2004, "The Seed of Apple's Innovation," Business WeekOnline.
[52] 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.
[53] 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.
[54] 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.
[55] Fischer, G., 2013, "Learning, Social Creativity, and Cultures of Participation," Learning and Collective Creativity: Activity-Theoretical and Sociocultural Studies, p. 198.
[56] Martin, M. W., 2006, "Moral creativity in science and engineering," Science and engineering ethics, 12(3), pp. 421-433.
[57] Nelson, B., and Yen, J., 2009, "Refined metrics for measuring ideation effectiveness," Design Studies, 30, pp. 737-743.
[58] Shah, J., Vargas-Hernandez, N., and Smith, S. M., 2003, "Metrics for Measuring Ideation Effectiveness," Design Studies, 24, pp. 111-124.

[59] Amabile, T., 1982, "Social psychology of creativity: A consensusual assessment technique," Journal of Personality and Social Psychology, 43, pp. 997-1013.
[60] 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.
[61] 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.
[62] 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.
[63] 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.
[64] 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.
[65] 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.
[66] Visser, W., 2006, The cognitive artifacts of designing.
[67] Sternberg, R. J., and Ben-Zeev, T., 2001, Complex Cognition, Oxford University Press, New York.
[68] EF, R., BA, N., and W., S., 2010, "The selection of creative ideas after individual idea generation: choosing between creativity and impact.," British Journal of Psychology, 101(1), pp. 47-68.
[69] Dyro, J., 2004, Clinical Engineering Handbook, Elsevier Academic Press, London, UK.
[70] Wickens, C. D., Lee, J. D., Liu, Y., and Gordon Becker, S. E., 2004, An Introduction to Human Factors Engineering, Pearson Prentice Hall, Upper Saddle River.
[71] Ward, T. B., Smith, S. M., and Finke, R. A., 1999, "Creative Cognition," Handbook of Creativity, R. J. Sternberg, ed., Cambridge University Press, New York, pp. 189-212.
[72] Weber, E. U., Böckenholt, U., Hilton, D. J., and Wallace, B., 1993, "Determinants of diagnostic hypothesis generation: effects of information, base rates, and experience," Journal of Experimental Psychology: Learning, Memory, and Cognition, 19(5), p. 1151.
[73] Pauker, S. G. K., R.I, 1992, "Clinical Problem-Solving: Risky Business," New England Journal of Medicine, 327(17), pp. 1241-1242.
[74] Pugh, S., 1991, Total Design, Addison-Wesley.
[75] Marsh, E. R., Slocum, A. H., and Otto, K. N., 1993, "Hierarchical decision making in machine design," MIT Precision Engineering Research Center.
[76] Pahl, G., and Beitz, W., 1984, Engineering Design, The Design Council, London.
[77] Sireli, Y., Kauffmann, P., and Ozan, E., 2007, "Integration of Kano's Model into QFD for Multiple Product Design," IEEE Transactions on Engineering Management, 54(2), pp. 380-390.
[78] 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.

[79] Gray, D. E., Brown, S., and James, M., 2010, "NUF Test," Gamestorming:, O'Reilly Media, Sebastopol, CA.
[80] Salonen, M., and Perttula, M., 2005, "Utilization of Concept Selection Methods: A Survey of Finnish Industry," Design Engineering Technical ConferencesLong Beach, California.
[81] Muller, G., Klever, D., Bjørnsen, H. H., and Pennotti, M., 2011, "Researching the application of Pugh Matrix in the sub-sea equipment industry by," CSER.
[82] Katsikopoulos, K. V., 2012, "Decision methods for design: insights from psychology," Journal of Mechanical Design, 134(8), p. 084504.
[83] Dong, A., Hill, A. W., and Agogino, A. M., 2004, "A document analysis method for characterizing design team performance," Journal of Mechanical Design, 126, p. 378.
[84] 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.
[85] 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.
[86] 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.
[87] Pedersen, T., Patwardhan, S., and Michelizzi, J., "WordNet:: Similarity: measuring the relatedness of concepts," Proc. Demonstration Papers at HLT-NAACL 2004, Association for Computational Linguistics, pp. 38-41.
[88] Miller, G. A., 1995, "WordNet: a lexical database for English," Communications of the ACM, 38(11), pp. 39-41.
[89] Pedersen, T., "Duluth: Measuring degrees of relational similarity with the gloss vector measure of semantic relatedness," Proc. Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, Association for Computational Linguistics, pp. 497-501.
[90] Amabile, T. M., 1983, "The social psychology of creativity: A componential conceptualization," Journal of personality and social psychology, 45(2), pp. 357-376.
[91] Rogers, E., 1995, "M.(1995). Diffusion of innovations," The Free Press, New York.
[92] Tversky, A., and Kahneman, D., 1974, "Judgment under uncertainty: Heuristics and biases," science, 185(4157), pp. 1124-1131.
[93] Wu, W., Luther, K., Pavel, A., Hartmann, B., Dow, S., and Agrawala, M., 2013, "CrowdCritter: Strategies for Crowdsourcing Visual Design Critique."
[94] 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.
[95] 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.
[96] 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.

[97] 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.
[98] Miller, S. R., Bailey, B. P., and Kirlik, A., 2014, "Exploring the Utility of Bayesian Truth Serum for Assessing Design Knowledge," Human-Computer Interaction.
[99] Kolb, P., 2008, "Disco: A multilingual database of distributionally similar words," Proceedings of KONVENS-2008, Berlin.
[100] 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.
[101] 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.
[102] Rogers, Y., Sharp, H., and Preece, J., 2012, Interaction Design - Beyond Human-Computer Interaction, 3rd Edition, Wiley\.
[103] Fallman, D., 2008, "The interaction design research triangle of design practice, design studies, and design exploration," Design Issues, 24(3), pp. 4-18.
[104] Onarheim, B., and Christensen, B. T., 2012, "Distributed idea screening in stage-gate development processes," Journal of Engineering Design, 23(9), pp. 660-673.
[105] Cooper, R. G., Edgett, S. J., and Kleinschmidt, E. J., 2002, "Optimizing the stage–gate process: what best-practice companies do," Research Technology Management, 45, pp. 21-27.
[106] 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.
[107] Ter Harr, S., Clausling, D., and Eppinger, S., 1993, "Integration of Quality Function Deployment in the Design Structure Matrix," Cambridge, MA.
[108] 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.
[109] 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.
[110] 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.
[111] 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.
[112] 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.
[113] 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.
[114] Sarkar, P., and Chakrabarti, A., 2011, "Assessing Design Creativity," Design Studies, 32, pp. 348-383.

[115] 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.
[116] Maher, M. L., and Fisher, D. H., "Using AI to evaluate creative designs," Proc. 2nd International Conference on Design Creativity, Glasgow, UK.
[117] Gero, J. S., 1990, "Design Prototypes: A knowledge representation schema for design," AI Magazine, pp. 22-36.
[118] 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.
[119] Cardoso, C., Goncalves, M., and Badke-Schaub, P., 2012, "Searching for Inspiration During Idea Generation: Pictures or Words?," International Design ConferenceCroatia, pp. 1831-1840.
[120] 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.
[121] Cross, N., 2004, "Expertise in design:an overview," Design Studies, 25(5), pp. 427-441.
[122] 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.
[123] 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.
[124] 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.
[125] 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.
[126] 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.
[127] 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.
[128] 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.
[129] 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.
[130] 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.
[131] 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.
[132] 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.

[133] 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.
[134] 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.
[135] 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.
[136] Nikander, J. B., Liikkanen, L. A., and Laakso, M., 2014, "The preference effect in design concept evaluation," Design Studies, 35(5), pp. 473-499.
[137] Von Hippel, E., 1986, "Lead users: a source of novel product concepts," Management science, 32(7), pp. 791-805.
[138] 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.
[139] 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.
[140] 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.
[141] Kolb, P., "Experiments on the difference between semantic similarity and relatedness," Proc. Proceedings of the 17th Nordic Conference on Computational Linguistics-NODALIDA'09.
[142] 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.
[143] Landis, J. R., and Koch, G. G., 1977, "The measurement of observer agreement for categorical data," biometrics, pp. 159-174.
[144] 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.
[145] 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.
[146] Chan, S., 1982, "Expert judgments made under uncertainty: Some evidence and suggestions.," Social Science Quarterly, 63, pp. 428-444.
[147] Anderson, J. R., 1987, "Skill acquisition: Compilation of weak-method problem situations," Psychological review, 94(2), p. 192.
[148] 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.
[149] Breen, B., 2004, "The 6 Myths of Creativity: A new Study Will Change How You Generate IDeas and Decide Who's Really Creative in Your Company," Fast Company, Fast Company & Inc.
[150] Cross, N., 1999, "Design research: A disciplined conversation," Design issues, pp. 5-10.
[151] Bharadwaj, S., and Menon, A., 2000, "Making innovation happen in organizations: individual creativity mechanisms, organizational creativity mechanisms or both?," Journal of product innovation management, 17(6), pp. 424-434.

[152] Laakso, M. L., and Liikkanen, L. A., 2012, "Dubious Role of Formal Creativity Techniques in Professional Design," International Conference on Design CreativityGlasgow, UK.
[153] Yilmaz, S., and Seifert, C. M., 2011, "Creativity through design heuristics: A case study of expert product design," Design Studies, 32(4), pp. 384-415.
[154] Nikander, J. B., Liikkanen, L. A., and Laakso, M., "Naturally emerging decision criteria in product concept evaluation," Proc. DS 75-7: Proceedings of the 19th International Conference on Engineering Design (ICED13), Design for Harmonies, Vol. 7: Human Behaviour in Design, Seoul, Korea, 19-22.08. 2013.
[155] Moreau, C. P., and Herd, K. B., 2010, "To Each His Own? How Comparisons with Others Influence Consumers’ Evaluations of Their Self‐Designed Products," Journal of Consumer Research, 36(5), pp. 806-819.
[156] Goncalo, J. A., Flynn, F. J., and Kim, S. H., 2010, "Are two narcissists better than one? The link between narcissism, perceived creativity, and creative performance," Personality and Social Psychology Bulletin, 36(11), pp. 1484-1495.
[157] Tauhid, S., and Okudan, G., "Fuzzy information axiom approach for design concept evaluation," Proc. International Conference on Engineering Design.
[158] Oladiran, M., Uziak, J., Eisenberg, M., and Scheffer, C., 2011, "Global engineering teams–a programme promoting teamwork in engineering design and manufacturing," European Journal of Engineering Education, 36(2), pp. 173-186.
[159] May, E., and Strong, D. S., 2006, "Is Engineering Education Delivering what Industry Requires," 3rd Canadian Design Engineering ConferenceToronto.
[160] Lingard, R., and Barkataki, S., "Teaching teamwork in engineering and computer science," Proc. Frontiers in Education Conference (FIE), 2011, IEEE, pp. F1C-1-F1C-5.
[161] Thiry, H., Laursen, S. L., and Hunter, A.-B., 2011, "What experiences help students become scientists?: A comparative study of research and other sources of personal and professional gains for STEM undergraduates," The Journal of Higher Education, 82(4), pp. 357-388.
[162] Bechtoldt, M. N., De Dreu, C. K., Nijstad, B. A., and Choi, H.-S., 2010, "Motivated information processing, social tuning, and group creativity," Journal of personality and social psychology, 99(4), p. 622.
[163] Simsarian, K. T., "Take it to the next stage: the roles of role playing in the design process," Proc. CHI'03 extended abstracts on Human factors in computing systems, ACM, pp. 1012-1013.
[164] Thomke, S., and Nimgade, A., 2000, IDEO product development, Harvard Business School Cambridge, MA.
[165] Kohn, N. W., Paulus, P. B., and Choi, Y., 2011, "Building on the Ideas of Others: An Examination of the Idea Combination Process," Journal of Experimental Social Psychology, 47(3), pp. 554-561.
[166] Larey, T. S., and Paulus, P. B., 1999, "Group preference and convergent tendencies in small groups: A content analysis of group brainstorming performance," Creativity Research Journal, 12(3), pp. 175-184.
[167] Naquin, C. E., and Tynan, R. O., 2003, "The team halo effect: why teams are not blamed for their failures," Journal of Applied Psychology, 88(2), p. 332.
[168] Briggs, K. C., 1976, Myers-Briggs type indicator, Consulting Psychologists Press Palo Alto, CA.

[169] Chang, T., and Chang, D., "The role of Myers-Briggs Type Indicator in electrical engineering education," Proc. Proceedings of the international conference on engineering education (ICEE 2000), Taipei.
[170] Osborn, A. F., 1953, "Applied imagination."
[171] Diehl, M., and Stroebe, W., 1987, "Productivity loss in brainstorming groups: Toward the solution of a riddle," Journal of personality and social psychology, 53(3), p. 497.
[172] Barki, H., and Pinsonneault, A., 2001, "Small Group Brainstorming and Idea Quality Is Electronic Brainstorming the Most Effective Approach?," Small Group Research, 32(2), pp. 158-205.
[173] Girotra, K., Terwiesch, C., and Ulrich, K. T., 2010, "Idea generation and the quality of the best idea," Management Science, 56(4), pp. 591-605.
[174] Rietzschel, E. F., Nijstad, B. A., and Stroebe, W., 2006, "Productivity is not enough: A comparison of interactive and nominal brainstorming groups on idea generation and selection," Journal of Experimental Social Psychology, 42(2), pp. 244-251.
[175] Sniezek, J. A., and Henry, R. A., 1989, "Accuracy and confidence in group judgment," Organizational behavior and human decision processes, 43(1), pp. 1-28.
[176] Garfield, M. J., Taylor, N. J., Dennis, A. R., and Satzinger, J. W., 2001, "Research report: modifying paradigms—individual differences, creativity techniques, and exposure to ideas in group idea generation," Information Systems Research, 12(3), pp. 322-333.
[177] Kerr, N. L., and Tindale, R. S., 2011, "Group-based forecasting?: A social psychological analysis," International Journal of Forecasting, 27(1), pp. 14-40.
[178] Charyton, C., 2014, Creative Engineering Design Assessment: Background, Directions, Manual, Scoring Guide and Uses, Springer.
[179] "ASME: Mission, Vision, and Strategic Priorities," https://http://www.asme.org/about-asme/who-we-are/mission-vision-and-strategic-focus.
[180] Luther, K., "Fast, Accurate, and Brilliant: Realizing the Potential of Crowdsourcing and Human Computation," Proc. CHI 2011 Workshop on Crowdsourcing and Human Computation, Vancouver, Canada.
[181] Luther, K., Caine, K., Ziegler, K., and Bruckman, A., "Why it works (when it works): success factors in online creative collaboration," Proc. Proceedings of the 16th ACM international conference on Supporting group work, ACM, pp. 1-10.
[182] Ferraris, C., and Carveth, R., "NASA and the Columbia Disaster: Decision-making by Groupthink," Proc. Proceedings of the 2003 Association for Business Communication Annual Convention, p. 12.
[183] Leikin, R., 2009, "Exploring mathematical creativity using multiple solution tasks," Creativity in mathematics and the education of gifted students, pp. 129-145.
[184] Hsieh, H. Y., 2014, "The Influence of the Designer's Expertise on Emotional Responses," Human Interface and the Management of Information. Information and Knowledge Design and Evaluation, Springer, pp. 572-582.
[185] Sternberg, R., 1999, Handbook of Creativity, Cambridge University Press, New York.
[186] Toh, C. A., Miller, S. R., and Kremer, G. E., 2013, "The Role of Personality and Team-based Product Dissection on Fixation Effects," Advances in Engineering Education, 3(4), pp. 1-23.
[187] 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).

[188] Li, Y., Wang, J., Li, X., and Zhao, W., 2007, "Design creativity in product innovation," The international journal of advanced manufacturing technology, 33(3-4), pp. 213-222.

[189] Nordstrom, K., and Korpelainen, P., 2011, "Creativity and inspiration for problem solving in engineering education," Teaching in Higher Education, 16(4), pp. 439-450.

Design Assessment Toolkit

Participation: *Individual Activity*

Overview:

You have recently generated and selected a few promising designs as a team. In the following activity, we will be introducing you to two methods for selecting early designs. The first method relies heavily on your individual intuition and experience to select the best designs to invest your team's time and efforts. The second method relies on you selecting words to describe each design and using these words to calculate numeric scores. By comparing these two methods, you will learn how to support your design selections using persuasive language and quantifiable data.

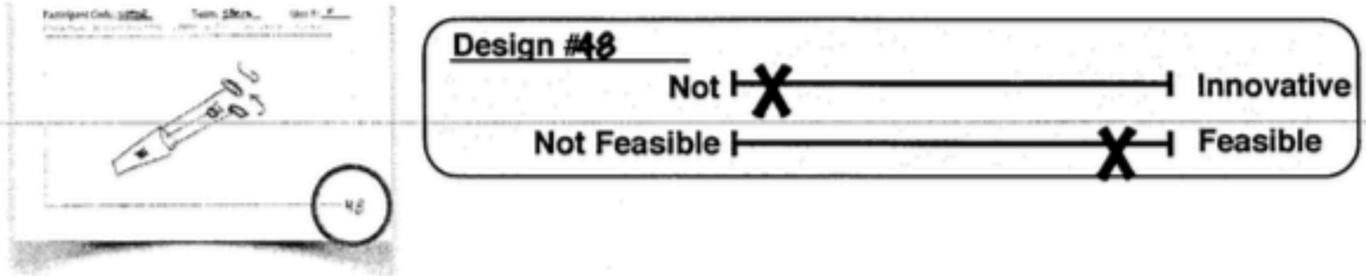
Instructions:

1.) Sit with your team and spread out all of the designs you are considering to move on in the design process in the middle of your table, sketch side up, so all team members can reach all designs.

2) Count how many concepts are in your team's pile and write that number here _____.

3.) Next, each team member should write their name and team name on the their ***DAT concept evaluation sheet*** (next page).

4.) Once this is complete, each team member should pull one design from the deck and individually evaluate the design for its innovativeness and feasibility by placing 'X's on the two lines of the ***DAT concept evaluation sheet (next page)****- see example below. Important: use the number in the bottom right hand corner of the design concept you are evaluating.*



5.) When you have finished evaluating your first design's innovativeness and feasibility, return the idea sketch side up in the middle of the table. Repeat step 4 until you have evaluated every design in the pile once. Double check that your total number of designs evaluated matches the number of total designs in step 2. Once complete please wait for instruction.

DAT *Individual Judgement*

Participant Code: Team:

Instructions:

Please place a X's on the lines below between the two extremes that most appropriately represents how innovative and feasible each of your team's designs are.

Design #: 42-example

Not Innovative |———————X———| Innovative

Not Feasible |——X———————————| Feasible

Design #: the number on the bottom right corner of each sketch.

Design #:

Not Innovative |——————————| Innovative

Not Feasible |——————————| Feasible

Design #:

Not Innovative |——————————| Innovative

Not Feasible |——————————| Feasible

Design #:

Not Innovative |——————————| Innovative

Not Feasible |——————————| Feasible

Design #:

Not Innovative |——————————| Innovative

Not Feasible |——————————| Feasible

Design #:

Not Innovative |——————————| Innovative

Not Feasible |——————————| Feasible

Design #:

Not Innovative |——————————| Innovative

Not Feasible |——————————| Feasible

Design #:

Not Innovative |——————————| Innovative

Not Feasible |——————————| Feasible

Design #:

Not Innovative |——————————| Innovative

Not Feasible |——————————| Feasible

DAT *Individual Judgement*

Participant Code: Team:

Instructions:
Please place a X's on the lines below between the two extremes that most appropriately represents how innovative and feasible each of your team's designs are.

Design #: 42-example

Not Innovative |————————X———| Innovative

Not Feasible |——X————————————| Feasible

Design #: the number on the bottom right corner of each sketch.

Design #: ____________

Not Innovative |————————————————| Innovative

Not Feasible |————————————————| Feasible

Design #: ____________

Not Innovative |————————————————| Innovative

Not Feasible |————————————————| Feasible

Design #: ____________

Not Innovative |————————————————| Innovative

Not Feasible |————————————————| Feasible

Design #: ____________

Not Innovative |————————————————| Innovative

Not Feasible |————————————————| Feasible

Design #: ____________

Not Innovative |————————————————| Innovative

Not Feasible |————————————————| Feasible

Design #: ____________

Not Innovative |————————————————| Innovative

Not Feasible |————————————————| Feasible

Once you have completed the ratings, please order your designs by their number from most effective to least appropriate solution.

Least									**Most**

DAT *Exercise Form*

Participant Code: Team:

Participant code: Last character of first name, 2 digits of day of birth, last 2 characters of birthplace
Design #: The number on the bottom right corner of each sketch.

Design #:

	Word 1:	**Word 2:**	**Word 3:**	**Word 4:**	**Word 5:**
W1					
W2					

Average Weights

+ Total

Design #:

	Word 1:	**Word 2:**	**Word 3:**	**Word 4:**	**Word 5:**
W1					
W2					

Average Weights

+ Total

Design #:

	Word 1:	**Word 2:**	**Word 3:**	**Word 4:**	**Word 5:**
W1					
W2					

Average Weights

+ Total

Design #:

	Word 1:	**Word 2:**	**Word 3:**	**Word 4:**	**Word 5:**
W1					
W2					

Average Weights

+ Total

DAT *Exercise Form*

Participant Code: Team:

Participant code: Last character of first name, 2 digits of day of birth, last 2 characters of birthplace
Design #: The number on the bottom right corner of each sketch.

Design #:

	Word 1:	**Word 2:**	**Word 3:**	**Word 4:**	**Word 5:**	**Average Weights**
W1						
W2						

+ **Total**

Design #:

	Word 1:	**Word 2:**	**Word 3:**	**Word 4:**	**Word 5:**	**Average Weights**
W1						
W2						

+ **Total**

Design #:

	Word 1:	**Word 2:**	**Word 3:**	**Word 4:**	**Word 5:**	**Average Weights**
W1						
W2						

+ **Total**

Design #:

	Word 1:	**Word 2:**	**Word 3:**	**Word 4:**	**Word 5:**	**Average Weights**
W1						
W2						

+ **Total**

DAT Word Selection

Your Name: Team:

Participant code: Last character of first name, 2 digits of day of birth, last 2 characters of birthplace
Design #: The number on the bottom right corner of each sketch.

Design #:

	Word 1:	Word 2:	Word 3:	Word 4:	Word 5:
W1					
W2					

Average Weights

+ **Total**

Design #:

	Word 1:	Word 2:	Word 3:	Word 4:	Word 5:
W1					
W2					

Average Weights

+ **Total**

Design #:

	Word 1:	Word 2:	Word 3:	Word 4:	Word 5:
W1					
W2					

Average Weights

+ **Total**

Once you have completed the ratings, please order your designs by their number from most effective to least appropriate solution.

Least									Most

DAT *What do you think?*

Participant Code: Team:

1.) How did your concept ratings differ between the two methods?

2.) Which rating method results (word selection or scale ratings) did you agree with more? Why?

3.) Did the word selection method alter the way you thought about your design ideas? Why or why not?

4.) What new insights did you gain from using the word selection method that you did not get using just the pure ratings?

DAT *What do you think? (continued)*

Participant Code: Team:

5.) Please complete the following chart by putting an 'X' in the appropriate box below.

This activity was more:	Rating Scale Method	Word Selection Method
useful		
fun		
efficient		
effective		
objective		
accurate		
insightful		

Provide rationale for your ratings below.

6.) If given the choice in the future, which activity would you use to effectively evaluate design concepts? Why?

Design Assessment Toolkit

Participation: *Team Activity*

Overview:

Now that you have completed the individual activity, the next activity will involve some teamwork. You will be reviewing each team member's design scores and calculating an overall score for each design. This activity will give you an opportunity to see the words that others have chosen and to think about the message each design has communicated.

Instructions:

1.) Briefly discuss within your team the words each of you has chosen.

2.) Next, review the ***DAT exercise team summary*** form. Have one person write the Design # (number on the bottom right of sketch) for each of your designs and your team's name on the the ***DAT exercise team summary***.

3.) Have each teammate take a turn and write their total design scores into the blanks under the corresponding Design # (number on the bottom right corner of sketch).

4.) Afterward, select one teammate to compute the average score of each design and write the calculated values in the final row of blanks corresponding to the Design #(number on the bottom right corner of sketch).

5.) Briefly discuss the final scores for each design.

6.) Finally, return to your individual ***Design Assessment Toolkit*** packet and complete the remaining questions individually.

Deliverables:

- A completed ***Design Assessment Toolkit*** packet from each teammate and one for the team's overall scores.

DAT *Exercise Team Summary*

Team:

Design #: the number on the bottom right corner of each sketch.

Design #				
Teammate 1				
Teammate 2				
Teammate 3				
Teammate 4				
Teammate 5				
Averages				

Design #				
Teammate 1				
Teammate 2				
Teammate 3				
Teammate 4				
Teammate 5				
Averages				

DAT *Exercise Team Summary*

Participant Code: Team:

Design #: the number on the bottom right corner of each sketch.

Design

Teammate 1				
Teammate 2				
Teammate 3				
Teammate 4				
Teammate 5				
Averages				

Design

Teammate 1				
Teammate 2				
Teammate 3				
Teammate 4				
Teammate 5				
Averages				

DAT *What does your team think?*

Rankings of your team's designs based on team averaged word scores.

Most									Least

1.) How do these ratings differ from your team's initial ratings formed in the discussion session (with the iPads / post-its)?

2.) Does your team agree with these new ratings? Why or why not?

3.) Based on all of the evaluation activities completed, which design, or designs, will you prototype?