The Pennsylvania State University

The Graduate School

VALIDATING CREATIVITY METRICS USING THE CREATIVITY METRIC EVALUATION FRAMEWORK

A Thesis in

Engineering Design

by

Sharath Kumar Ramachandran

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2019

The thesis of Sharath Kumar Ramachandran was reviewed and approved* by the following:

Scarlett R. Miller
Associate Professor, Engineering Design and Industrial Engineering
Thesis Advisor

Samuel T. Hunter
Associate Professor, Psychology

Sven G. Bilén
Professor of Engineering Design, Electrical Engineering, and Aerospace Engineering
Head, School of Engineering Design, Technology, and Professional Programs

*Signatures are on file in the Graduate School

ABSTRACT

The ability to reliably evaluate creativity of ideas generated in the early phases of design is an important and difficult challenge for design researchers. While a variety of metrics have been developed for these purposes, there is yet to be a standardized framework for evaluating the validity of existing or new metrics in our field. Establishing such a framework is important because prior research has shown that using different creativity metrics and yield vastly different, and even sometimes negatively correlated results. These findings lead to some important research questions: "Are these metrics measuring what we're intend to measure?" and "What are the implications of our measurements if the validity of the metrics currently used is untested?"

Thus, the purpose of this thesis was to create a structured framework for testing the validity of engineering creativity metrics. This was achieved by the creation of the CreAtivity Metric Evaluation fRamework (CAMbER) and through a demonstration of its utility through a case study of design variety. Specifically, the CAMbER framework developed as part of this thesis work includes a stepwise methodology aimed at identifying a creativity construct to examine, establishing a ground truth, measuring the validity of the metric under the examination, and identifying how the validity of the metric is impacted by modifications in the metrics computation. The variety metric; generally used to measure the spread of ideas within a sample set generated, was used to demonstrate how the CAMbER framework could be deployed in engineering design research. The results of this thesis will not only help study and validate existing engineering creativity metrics, but also provide a guiding light to ensure the creation of reliable creativity metrics in the future.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Firstly, I would like to thank my advisor, Dr. Scarlett Miller, for providing me with the opportunity to conduct the following research under her guidance. Your constant support and encouragement made the completion of this thesis possible. Your expertise in the area and patience with me when I made mistakes, helped make this work possible. I thank you for providing me with the opportunity to present at international conferences, write technical papers and win poster competitions abroad. Thank you, for welcoming me to be a part of the BRITE lab family, and contribute towards these exciting research projects.

I would like to thank Dr. Mark Fuge and Faez Ahmed, from the Department of Mechanical Engineering, University of Maryland, our collaborators, and instrumental contributors in our creativity research projects. This thesis would not have been possible without your efforts and support.

Next, I would like to thank Dr. Samuel T. Hunter for his inputs and guidance in the field of psychological research, and providing valuable feedback on the papers. Thank you, Dr. Elizabeth Starkey for your guidance and collaborative work in our projects.

I would like to thank my BRITE lab family, Katie, Rohan, Hong-En, Liz, Xuan and Mo for working and celebrating together, as colleagues and friends. Thank you, Zibing (Janice) and Siddarth for your efforts and assistance in the analysis we ran together. Thank you, Terri and Katie, for your constant support, and for checking on me when I needed it most.

I would like to thank my family – Daddy, Mummy and Anna (Praveen), for your incredible support and belief in me during my master's program and my time away from home. You have supported me, blessed me, and guided me all my life. It's your hard work and sacrifices that have helped me realize my professional and academic goals. Finally, I would like to thank my friends, and everyone back home for their support and belief in me over these two years.

This research was funded by the National Science Foundation under Grant No.1728086. Any opinions, findings, and conclusions expressed in this thesis are those of the authors and do not necessarily reflect the views of the NSF.

CHAPTER 1

INTRODUCTION

Creativity is a crucial part of the engineering design process as it enables innovation [1] and is the quintessence of new businesses [2]. As such, researchers have devoted substantial effort to developing and testing methods for supporting creative idea development through idea generation methods (see for example [3-10]). As the methods to develop these solutions increase, the necessity to develop rigorous techniques for evaluating the success of these methods also increases, because the success of research in this field hinges on the quality of the measurements used for interpreting research findings [11]. While a plethora of methods exist to measure creativity in engineering design research, (see for example [12-19]), engineering creativity metrics however, have been criticized for their lack of generalizability across domains [20], the subjectivity of the measurements [21], and the vagueness of the measurement methods [22]. This is important because we can only assess and control what can be accurately measured [23]. As such, having a standardized methodology for validating the constructs under examination is crucial to the success of creativity research [24].

Construct validation focuses on the agreement between theoretical and operational definitions of an attribute under observation [25]. In other words, it is the extent to which a process measures the characteristic it is supposed to measure [26, 27]. It is one of the three principle methodologies (content-related, context-related and construct-related validity) outlined by the American Psychological Association (APA) [28]. However, the APA has acknowledged the

difficulty in measuring abstract constructs like creativity due to its high level of abstraction. They also stress that this greater level of abstraction also increases the need to validate the construct, so as to reduce the interpretation on the user's end [29, 30]. While there are a wide variety of means available to assess construct-validity, including analyses of internal consistencies [31-33], correlations of new measures [34], and expert assessments relevant to the construct under examination [29, 34], there has been limited construct validation performed in the engineering creativity literature.

One factor that has limited construct validation in the engineering creativity literature is the breadth of definitions that exist for the word 'creativity'. In fact, a recent meta-analysis identified more than 160 definitions, which makes measuring the nebulous construct of creativity difficult [19]. For instance, in the social sciences, Amabile (1983) defines creativity as "*the process by which something judged (to be creative) is produced*" (pg. 1001 [35]) while Torrance (1966) scored creativity based on fluency, flexibility, inventive level and elaboration (pg. 245 [36]). On the other hand, some authors provide no definition of the word creativity at all and instead opt to avoid using the term 'creativity' in the description of their metrics due to the "*difficulty in defining this term (and agreeing on its meaning)*" (pg. 116 [16]). Instead, these authors state their metrics instead measure 'ideation effectiveness' rather than creativity. As such, having a clearly articulated definition of the construct under examination is the first aspect of creating and validating a metric as a lack of a firm or accurate foundation can deplete the validity of a metric [37].

Another factor that has hindered construct validation is a lack of a ground truth against with which metric performance can be compared. Without a ground truth, validating creativity metrics is difficult because measuring validity is often associated with the accuracy of the measurement [38-40], or the ability to produce results closest to the actual phenomena (ground truth) of the measured artifact [40]. Because there is no definite paradigm to directly measure

creativity (e.g. establish a ground truth), indirect measures are often used [25, 41] to validate these metrics by comparing results against existing creativity metrics (see for example [15, 19, 42]). The problem with this approach is that since there is no ground truth in creativity research, if the results are contradictory, as was the case in the research carried out by Lopez, Zheng and Miller [43]), it is not clear which metric, if any, is more valid. This example brings up a very important and yet rarely discussed problem with existing engineering creativity metrics; existing engineering creativity measures often lack construct validation, meaning we do not know if, or to what effect, these metrics are actually measuring creativity. Thus, establishing a method for identifying or reporting ground truth is essential for metric validation.

Although creativity metrics have been widely used in the engineering literature for the past two decades, there has yet to be a standardized process for validating the accuracy of these creativity metrics or establishing a ground truth. Thus, the purpose of this thesis was to present a first step towards this goal by developing the CreAtivity Metric Evaluation fRamework (CAMbER), demonstrate its use through a case study of the design variety metric, and indentify how metrics may be refined and retested through this framework. The instrument and the methodology adopted in this thesis can be used to validate both existing engineering creativity metrics and forthcoming metrics in order to improve the reliability of creativity instruments used in the field.

1.1 Research Objectives and Significance

While a wide wealth of creativity metrics has been developed over the past three decades, researchers have failed to successfully provide validation for these metrics. In response to this, the current thesis was developed to be a first step at filling this research void. Specifically, the current

thesis focused on two main objectives: 1) to provide a framework which can be used by researchers to test the validity creativity metrics in engineering design research, and 2) to identify methods that can be used to modify metric calculations to improve metric validity. The remainder of this chapter discusses the contributions of this thesis and provides an outline of what this thesis work entails.

1.2 Expected Contributions

This thesis develops a theoretical framework that can be used to calibrate existing creativity metrics and metrics that are yet to be developed. The first part of the thesis, provides a step-wise analogy, the framework elaborates on a methodology to validate metrics by reiterating the importance of what is being measured. The opportunity to refine the metrics and compare the results to the ground truth provides a cyclic approach to ensure accurate measurement of creative characteristics in engineering design. The overarching purpose of this work is to provide methods to validate and refine metrics used to quantify constructs of creativity in the field of design engineering.

1.3 Document Outline

The remainder of this thesis elaborates on the metrics elected, methodologies used, findings, and the design implications of the results in detail. Specifically, Chapter 2 presents the CreAtivity Metric Evaluation fRamework (CAMbER) guiding the user through a metric validation framework using a 5-step methodology. Chapter 3 uses the Design Variety metric as an example to test for its validity using the CAMbER framework. Chapter 4 presents a methodology for using

human raters to modify creativity ratings. Finally, Chapter 5 presents the conclusion and implications of this work in the field of creativity research.

CHAPTER 2

The CreAtivity Metric Evaluation fRamework (CAMbER)

While the last chapter highlighted the key contributions of the thesis and an outline of the research work to follow, the current chapter shifts the focus to providing a details and theoretical foundation for the CreAtivity Metric Evaluation fRamework (CAMbER). Specifically, Figure 1 demonstrates the five-step framework used to evaluate the validity of creativity metrics in engineering design research. These steps include: identifying a construct for review (Step 1), selecting the type of metric validation (Step 2), establishing and reporting the ground truth (Step 3), evaluating metric validity (Step 4), and refining the metric based on findings (Step 5). The remainder of this chapter serves to provide a foundational basis for these steps. Following this discussion, a case study is presented in the following chapter using this framework to explore the accuracy of the design variety creativity metric.

Step 1: Identify Construct for Review

As Figure 1 demonstrates, Step 1 of the CAMbER framework involves identifying the construct, or the attribute under observation, for review. This is particularly vital in the CAMbER framework because of the widely varying definitions of creativity [19]. In other words, what is being measured must be identified before it can be determined how well a metric is measuring what it claims to be measuring [29].

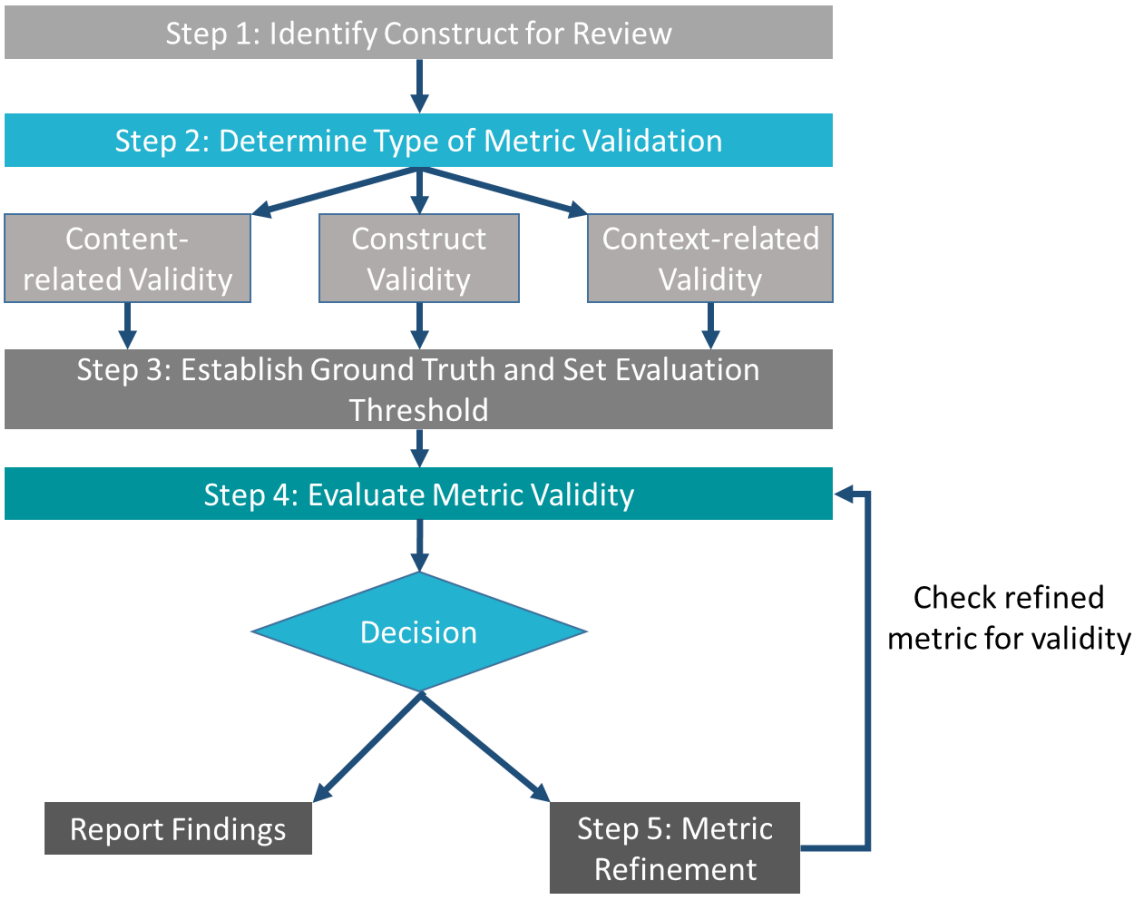

Figure 1: Metric Validation and Refinement Framework

An example of a construct definition is the length of the meter, which has been defined as one-millionth of the distance between the Equator and the North Pole by the French Academy of Sciences in 1791 [44]. However, over time it has been identified that this definition is not as accurate because the dimensions of the earth are transient. Because of this, the definition of this construct has evolved over time to be something more robust; today, the definition of a meter is "*the length of path traveled by light in vacuum in a time interval of 1/299,792,458th of a second*" [45]. This example not only illustrates the importance of having a clear construct definition, but it also demonstrates that a constructs definition should be revisited over time. As such, it is essential

that it is reported when we describe a metric for validation. This is particularly important with creativity research due to the abstract nature of creativity. These types of abstract constructs are typically based on subjective perception rather than experiential affirmation [25]. Thus, establishing the construct being evaluated as part of the CAMbER framework is essential in order to ensure valid measures.

In order to assess construct validity, Validity differs from validation through the fact that validity refers to a quality of something, while validation refers to the actions taken to evaluate these aspects of validity [46]. Before one can begin to evaluate the validity of a creativity metric, it is first important to define the construct for which you are evaluating. “Construct” is a term used in the social sciences to describe the idea or theory under observation, typically based on subjective perception rather than experiential affirmation [25]. Construct validity focuses on the agreement between theoretical and operational definitions of an attribute under observation [25]. The extent to which an operationalization measures the characteristic it is supposed to measure is essentially construct validity [26, 27]. Construct validation is involved whenever a test is elucidated as an estimate of a quality or an attribute that is not operationally defined [25]. Having a clear definition of the construct is important because the lesser the subjectivity in a measurement, the less likely there will be variance while the same attribute is measured by different judges using a given method. Construct validation occurs when a metric or instrument reflects a construct, with a meaning attached to it [25].

Step 2: Determine Type of Metric Validation

Once the construct has been identified, the next step is to identify what type of validation the researcher is trying to achieve. Validity is viewed as the degree to which a measure reflects what is actually being measured. The analytical technique of accumulating and appraising crucial data under examination is referred to as validation [29]. Questions that need to be answered by any metric are what is it measuring? And how well is it measuring what it claims to be measuring? [29]. In other words, the user of the metric must first understand why the metric is being used and what the results represent.

While there are numerous strategies to evaluate validity, the American Psychological Association standards [47] use three principal methodologies: content-related evidence [48], criterion-related evidence [49], and construct-related evidence [34]. Content-related validity involves making inferences from sampling a set of observations from a universe of information. The judgement in content-related evidence depends on the representativeness of the sampling and relies heavily on expert assessment [48, 50-53]. Much of the procedural and judgmental evidence in content-related validity stem from test development and procedures to derive test scores [52]. For example, in civil rights legislation, new types of cases are often encountered which lack expert opinions on the matter. In such situations, content-related evidence is generally adopted. A panel of members knowledgeable about the job are selected to independently answer a number of items that constitute a "shelf test" [52]. The test question asked for each of the item is as follows:

Is the skill (or knowledge) measured by this item

- *Essential*
- *Useful but not essential, or*

- *Not necessary*

to the performance of the job? (pg.567 [52])

If there is any discrepancy in the answers of the panel for each of the items, the level of knowledge or skill in the domain of each panelist is used to make judgement on each item [52].

On the other hand, construct-related validity provides a logical foundation that establishes the underlying definition of what the construct under observation represents. Though construct-related validity and content-related validity are similar, content-related validity can serve as a branch of construct-related validity [54]. Finally, criterion-related validity is used when attributes can be predicted based on a different outcome that measures particular differences [29]. An example that explains criterion-related validity is that the ability of students' performance in graduate schools is determined by the Graduate Record Examination (GRE) or that levels of anxiety are used to predict students' performances in exams [55]. Neither content-related validity nor criterion-related validity have the answer as to how these predictions are made. They simply provide correlations that are proven to be useful for a given sampling size. Of these three types of validations in the study of creativity, the most often utilized method for metric validation is construct-validation. TTCT was developed in the 1960s as one of the early attempts of studying creativity [36]. Several longitudinal studies have attempted to study the construct validity of TTCT [56-59]. A recent example where researchers have studied construct validity in the field of creativity is the testing of TTCT (Torrance Test of Creative Thinking) [60]. Fluency, flexibility, originality and elaboration in figurative and verbal form were used as cognitive processes to measure creativity. A factor analysis of principal components was applied to test the ideas for the TTCT dimensions. The consistency of the scores were tested to measure if they reflected the theory that creativity could be defined by cognitive dimensions [60].

Step 3: Establish a ground truth

Once a construct has been defined and the type of validity has been established, the next step is to determine a "ground truth" to use as a reference. Developing a ground truth for physical quantities like mass is commonplace; for example, the International Prototype Kilogram (IPK)—a 90% Platinum, 10% Iridium cylinder (with a height and diameter of 39mm [61]) stored at the International Bureau of Weights and Measures—acts as the official reference for one kilogram [62]. When a new scale is created, it can be calibrated to the kilogram that exists as a ground truth measure. However, establishing such a ground truth is not as straightforward in the social sciences. Creative expression varies from person to person and the number of cognitive, socio-effective and environmental variables that could interfere with the creative process is very large [63]. In order to overcome these difficulties, researchers have applied numerous strategies including: analyses of internal consistencies, correlations of new measures, and expert assessments relevant to the construct under examination [29, 34].

In statistical data collection, Brackstone [64] identified accuracy, relevance, timeliness, accessibility, interpretability and coherence as salient dimensions of the quality of data. In addition, several other studies have identified the quality of accuracy of data measured against the true value, and suggested its importance in the credibility of information collected and analyzed [65-68]. Establishing the ground truth and validating the constructs for every measurement technique in order to obtain acceptable results (close to the true value of what is being measured) is paramount in creativity research like any other field of science [68].

In the social sciences, the most commonly used, albeit imperfect, method for establishing a ground truth for creativity has been through subjective measures. For example, the Consensual Assessment Technique (CAT) put forth by Amabile [35, 69, 70], relies on the simple idea that an

artifact is creative only to the extent to which experts in the area agree, independently, that it is creative. In the CAT, raters are required to utilize internal heuristics and prior experience to rate aspects such as drawing abilities, elegance, creativity, usefulness and uniqueness on a 6 or 7-point Likert scale. This method has been widely integrated throughout social science research because creativity is seen as a complex construct which is inherently subjective. Importantly, these metrics have been shown to have high validity (e.g., [71, 72]). As such, the consensus among social scientists is that these subjective methods represent the best means for assessing creativity of a product or outcome. Amabile mentions that *"By definition, inter-judge reliability in this method is equivalent to construct validity; if appropriate judges independently agree that a given product is highly creative, then it can and must be accepted as such"* (pg.1002 [35]).

Apart from CAT, there has been prior work in the field of creativity research to measure accuracy of new and existing scales utilizing the judgement from experienced designers as the ground truth [17, 19, 73, 74]. Knowledge in a particular domain built over time evaluating similar designs and ideas provides as a measure to assess attributes of creativity [19]. This provides a basis that if a scale tends to agree with human expert raters (while the expert raters independently agree with each other), the scale can be deemed acceptable [35, 71]. Utilizing experienced raters, attributes of creativity have been assessed. The rankings have then been processed (Average ranking, Spearman's rank correlation, etc.) to obtain the baseline against which the accuracy of new and existing metrics have been tested [19, 73]. Relative rankings have been used to measure accuracy when ratings systems are evaluated on different scales [19].

The validity of these ratings depends on the premise that there are expert human judges who are consistent with their creativity ratings, and provide high inter-rater reliability. This provides a benchmark to assess creativity. Hence, the agreement of expert judges is can been seen as an argument for objectivity of the method while ensuring large amounts of responses as received

from multiple raters [35, 75-77]. Agreement of human raters has also been used to study word associations with creativity by Van Der Velde *et al*. [78] in their work with a semantic mapping of words related with creativity. Other evidence supporting the use of human raters in the use of scales used to evaluate creativity include Creativity Support Index [79] and the Creative Product Semantic Scale [80]. Human judgements are used as starting points as they precisely portray the intuitive concept of creativity [81]. Well established scales have proven to be competent in projecting the ground truth through subjective assessment by experts when they independently agree that an artifact or an idea is creative [35, 71, 72]. Hence, expert opinions on data that has been sampled to be representative can be used to measure accuracy of scales [72]. In the lens of creativity, such well-established and tested scales can be used to verify the accuracy of existing and new scales being developed.

In the field of creativity, human raters with experience/expertise are often used to benchmark attributes such as novelty and variety. For example, a study conducted by Chakrabarti and Khadilkar utilized 13 designers with backgrounds in engineering and architecture as human experts to arrive at the ground truth novelty of ideas that aimed at creating computer mice [17]. The SAPPhIRE (State-Action-Part-Phenomenon-Input-oRgan-Effect) model proposed by Sarkar and Chakrabarti in order to evaluate a novelty metric utilizes the knowledge of experienced designers who can serve as experts. The designers are experienced in regularly judging concepts and ideas in patent offices and design firms in similar domains. In absence of physical prototypes or fundamental constants to guide ground truth creation, human experts are selected to arrive at the ground truth [19]. Recently, Jagtap [73] worked at refining the SAPPhIRE model [19] to asses measure design novelty. 18 experienced designers were used to assess product novelty and these ratings were considered to be the ground truth. The modifications to the scale were later ascertained by comparing the existing scale and the new scale to the ratings provided by the 18 experienced

designers [73]. The designers were required to rank three sets of products for novelty within each set. The average rankings provided by the 18 designers were utilized as a benchmark against which the scales were later compared using Spearman's rank correlations [73]. These methods suggested above utilize human intuition based on experience in relevant fields to determine the ground truth [17, 19, 73, 74]. The next step discusses how the ground truth is utilized to test the accuracy of the metrics.

Step 4: Evaluate Metric Validity

Once a ground truth has been established, one can begin to evaluate the validity of the metric by first selecting the appropriate method for such comparison, and then selecting the appropriate threshold for said method. This step is very important to determine if the user need to move to step 5 to refine the metric or if the metric has produced substantially acceptable metrics that can be reported and used.

The most common validity measure in psychological research is the measure of internal consistencies. As internal consistencies are readily calculable from a single test administration, Hogan, Benjamin, and Brezinski [82] discovered that "*about 75% of reported reliability estimates in the Directory of Unpublished Experimental Mental Measures (published by the American Psychological Association [APA]) were internal consistency estimates.*" (pg. 177 [83]). The internal consistency scales relate to the degree to which items of a test align with each other to measure a similar construct. Internal consistencies have been vastly utilized in research to measure reliability in creativity [84-87]. Pearson correlation and Cronbach's Alpha coefficients have been utilized to measure internal consistencies to ensure each item and all items together measure the same attributes [88].

As discussed earlier, according to the methods prescribed by the American Psychological Association, three major types of validity are recommended based on the type of measurement in a given case; content-related validity, context-related validity and construct validity [25, 34, 47-49]. Construct validity is often chosen for concepts like creativity when the test or tool has to connect the theoretical and operational definitions of the attribute under measurement [25-27].

One of the examples of evaluating construct validity is as follows. The California Psychological Inventory Creativity Scale/Test (CPI-CT) [89, 90] and the Myers-Briggs Type Indicator Creativity Index (MBTI-CI) [91, 92] were used as personality tests to measure the construct of creativity as an individual difference variable. The Kirton Adaption-Innovation (KAI) inventory was proposed to measure the style of problem solving and creativity [93]. This scale is widely used in academia, business, psychology and sociology to locate people on the scale and interplay adaptors and innovators in organizational settings to solve problems more effectively [94]. In an experiment conducted by Fleenor and Taylor [95], three self-reported measures were evaluated to measure the nature and extent of relationships between them. It is to be noted that the CPI and MBTI measure levels of creativity while KAI measure the style of creativity. The CPI-CT is a 42-item scale that aims to measure the extent of creativity in an individual [90]. The MBTI-CI is a scale based on personality assessment research, where individuals judged to be creative were found to score higher on the MBTI-CI ranging between a high of 365.4 to a low of 221.1 [91]. The KAI inventory is a 33-item self-reported inventory used to measure cognitive styles in people abilities to make decisions and solve problems [94]. Correlation analysis was run between the three scales to analyze effects of creativity style and levels of creativity. This is an example of construct validation, where tested scales can be utilized to measure the accuracy of new or existing scales that aim to measure the same construct.

While evaluating any construct or internal validity, the extent to which the measure attribute reflects ground truth is important. Landis and Koch [96] tried to propose a method to evaluate multivariate categorical data from studies where observer reliability studies are conducted. The study utilized the data from the diagnosis of multiple sclerosis cases reported in a study conducted by Westlund and Kurland [97] in Winnipeg. After the examinations, each neurologist was required to asses all the examinations without looking at earlier summaries and categorize the patients as certain-, probable-, possible-, doubtful multiple sclerosis. The agreement between the diagnosticians was calculated and a Kappa statistic was calculated [96]. Landis and Koch arrived at meaningful representations to the Kappa statistic and the strength of agreement and they were as follows (Table 1):

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

Table 1: Strength of agreement against Kappa Statistic as proposed by Landis and Koch

There have been several examples of studies in the domain of creativity that utilize percentage agreement to understand the reliability between judges. In a study conducted by Fontenot [98], percentage agreement has been utilized to evaluate reliability and agreement between judges in the effects of creativity training and creative problem finding in business settings. Ward [99] measured creativity in young children using measures of divergent thinking. Evaluating agreement between judges was carried out using percentage agreement indices between their

responses [99]. In an evaluation of dance creativity, originality, flexibility and fluency as prescribed by Guilford and Hoepfner [100] were utilized to evaluate performance creativity. Percentage agreement has been utilized to evaluate the agreement between raters [101]. Percentage agreement is a powerful tool that not only helps understanding reliability between judges, but also to evaluate metrics that attempt to pursue the same goal.

Step 5: Metric Refinement

If Step 4 reveals low metric validity at the chosen threshold of a new metric, then the metric should undergo refinement and retesting. Metric refinement is crucial step in the system and completes the loop ensuring the measurements closest to the uncovered ground truth are achieved [29]. Refinement of a metric can include reviewing the data collected to eliminate any confounding effects, modifying weights or principles involved in the calculation of the metric, or reviewing definitions of the aspects of creativity under observation [102]. For example, recent research published by Jagtap [73] identified a method to refine the design novelty metric developed by Sarkar and Chakrabarti's novelty assessment method [19]. Specifically, Jagtap brought to light perceived deficiencies in the SAPPhIRE (State-Action-Part-Phenomenon-Input-oRgan-Effect) model [19]. Factors including differences in inputs and parts, differences in physical effect/phenomenon and considering inputs and structures to determine the novelty of the product are elaborated [73, 102]. The refined method was evaluated using intuitive assessment of novelty by experienced designers, who serve as experts in the field. Spearman's rank correlations were utilized to analyze the agreement between the old metric and the refined metric against human ratings. It was determined that there was a stronger agreement between the expert human-raters and the refined metric in comparison with Sarkar and Chakrabarti's original metric [73].

With progress in research, we are always at a constant motion to refine existing metrics to produce more accurate results [103-105]. There are examples in research where results from existing metrics are used to evaluate new metrics. In these cases, the ground truth is established using existing metrics and modifications to the metrics, or new metrics are compared to the original metrics and conclusions are drawn on how the modifications improve the method in which an attribute is measured, hence attempting to create a new ground truth [15]. The SVS novelty metric uses genealogical trees to differentiate ideas at each level of abstraction [16]. The metric developed by Vargas-Hernandez, N., Okudan, G., & Schmidt, L [106] aimed at improving the effectiveness of the SVS metric by merging genealogical trees, and in theory created a more accurate ground truth [106]. Another metric developed by Johnson *et al.* [15] proposed a metric that added another branch in the genealogical tree and in essence would (1) make meaningful measurements handling boundary cases, (2) support changes to the dataset and (3) support abstract responses. The new metric broadened the SVS metric similar to the Vargas-Hernandez metric but included the entire tree instead of a subset of the data. This was an attempt by Johnson *et al.* [15] to create a refined metric to measure design novelty.

Once a metric has been refined, it is important that the user returns to step 4 and retests the validity of the scale. If upon rigorous refinement and re-testing, the metric is still unable to produce values resembling the ground truth evaluation, the researchers will need to reconsider either the overall validity of the proposed method, or the validity of the ground truth measurement. Either way, the findings should be reported within the limitations of the study identifying key opportunities for future work.

CHAPTER 3

A CASE STUDY OF CAMbER DEPLOYMENT

While Chapter 2 outlines the theoretical foundation for the CAMbER framework, the current chapter turns our focus to the demonstration of its utility through a case study for the creativity metric *design variety.* Data for the case study was gathered from a previous experiment which included a subset of 10 ideas from the 934 ideas generated in an experiment conducted by Starkey, Hunter and Miller [107]. In this study, the participants were asked to generate ideas for a "novel and efficient milk frother."

Step 1: Identifying the Construct for Review

The design variety metric presented by Nelson et al. [108] (refined from the variety metric provided by SVS [16]) was selected as the construct to validate for the current case study. Design variety has been defined as how "explored the solution space" is during the idea generation process (pg. 117, [16]). The variety measure is used to delineate the extent of dissimilarity of ideas within a given set and is used to counterbalance the quantity measure in design studies [106]. This is because increases in the fluency of ideas must also proportionally increase the spread of the ideas [109]. This is important because research has shown there is a higher possibility of solving a design problem when a more discrete set of ideas is produced in the initial stages of the design process [110]. In addition, this type of divergent thinking has been shown to directly translate to the production of successful design solutions. The variety metrics commonly adopted in the engineering literature [16] breaks design variety into four hierarchical branches: the physical

principle, followed by the working principle, embodiment and detail levels. These levels were developed to ensure that "separation at higher levels will always score a greater total" of variety score (pg. 126 [16]). In other words, ideas that differ in physical principle are thought to diverge by a greater extent in the design space compared to ideas that only differ by minor details [16, 108].

The first step towards deploying these metrics was to define the physical, working, embodiment, and detail levels of the tree [16]. Each of the levels of the hierarchical tree were defined as follows for the milk frother problem: Physical principle: the form of motion that generates the frothing (ex: spinning, tumbling, air pressure, chemical reaction, etc.), Working principle: the type of energy source used to facilitate the frothing (ex: human-powered, electric power, battery power, gravity, etc.) and Embodiment: the physical representation of the idea (ex: looks like a bicycle, jet turbine, blender, etc.) [111]. The ideas generated and used in this study did not produce a detail level measure (SVS [16] articulates that in idea sets where extensive detail is not provided, the detail level can be disregarded in the construction of the tree and calculation of the metric). Figure 2 demonstrates how these principles were used to develop a genealogy tree for an example set of 6 ideas from the milk frother design problem.

Once the levels were defined, the computation of the variety metrics was completed according to the methods outlined in Nelson et al. [108]. These metrics differ in two distinct ways. First, while both methods seek to reward ideas that have more variety at higher abstraction levels by providing higher weights to more abstract principle, they differ in their assignment of weights. For example, SVS [16] assigned 10, 6, 3 and 1, for tree physical, working, embodiment, and detail respectively while Nelson et al. [108] for utilized 10, 5, 2, and 1 for the metric. Second, Nelson et al. [108] modified the computation to avoid double counting ideas. Figure 2 identifies how these two differences result in variations in the variety score achieved. In this example, as expected, the score computed through SVS [16] (Variety score = [(10*4) + (6*5) + (3*5)]/5 =17) was higher

than the score computed through the Nelson [108] metric (Variety score = [(10*3) + (5*1) + (2*0)]/4 = 8.75) as shown in the Figure 2.

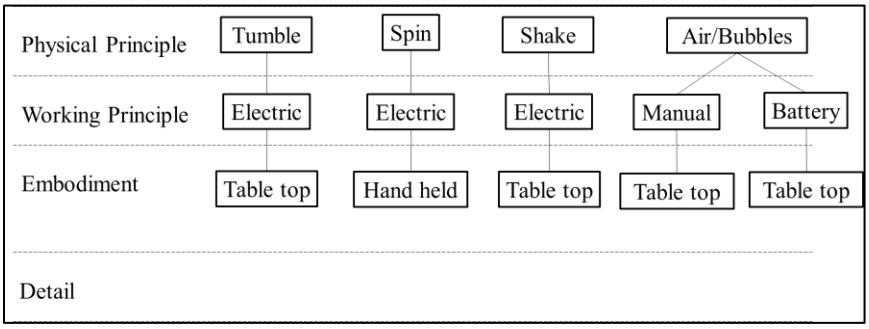

Figure 2: Genealogical tree with method of frothing as the physical principle and type of power used as the working principle. SVS variety score = 17, Nelson variety score = 8.75

Step 2: Determine Type of Metric Validation

As discussed in the previous step, the metric selected to validate was the variety metric developed by SVS [16] and later refined by Nelson et al. [14]. The attribute that the metric aims to measure is design space exploration. The ability to produce many possible solutions to a given problem stems from divergent thinking [112, 113]. Creativity measured through divergent thinking tasks is generally considered a domain construct [63]. Cronbach and Meehl [25] state that "Construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured." (pg.2). Though design variety has been researched over several decades, creativity and in specific design variety does not have a single definition that is acceptable based on the attribute measure. This is one of the reasons why researchers who have compared metrics to each other or metrics to human expert raters, have found varied correlations between them [14, 71, 73, 102]. Hence, for the measurement of design variety the type of metric validation was selected to be construct validity.

Step 3: Establishing a ground truth

Once the construct has been defined and computed, the next step is to establish a ground truth dataset for the calculation of design variety. For the case study, the first step in this process was to identify the best means of assessing the ground truth for design variety. For the purpose of this case study, expert human raters were utilized, as experts in relevant domains have been known to provide consistent results in assessing creativity [79-81]. Once the method for establishing the ground truth was set, the next step was to identify a ground truth set of sketches for which the raters would evaluate design variety. This is important because there were 1238 ideas in total to be evaluated.

Due to the limited time and availability of raters to assess all those ideas manually, 10 design sketches studied in Ahmed et al [39] and selected from [107] which were put into 210 sets of six sketches for comparison by human raters (10C6 = 210 i.e. total possibilities of 6 sketch sets from a pool of 10 sketches). Six ideas were selected from each set because this was the median number sketches made by participant's in the previous study [107]. Out of these 210 sets, 21,945 paired comparisons were possible to test. However, the limited time and availability of the rater, and the fatigue associated with rating this many comparisons was infeasible.

In order to meaningfully reduce the number of pairs rated, the 210 sets from the highest to lowest variety set were rank ordered using a pairwise distance metric. This metric was derived from an idea map research study examined by Faez et al. [114]. In this study, each participant was provided with the same 10 sketches and was asked to pin them on a canvas, such that the relative distances were proportional to their similarity. Sketches closer to each other on the canvas were more similar than sketches that were further apart. The pairwise distance metric was used to compute similarity based on distances between the sketches for each of the participants. Using this process, the absolute rank difference between the two items for each comparison was calculated. A

small rank difference implied that the two sets had similar variety, while a large rank difference implied that the metric is confident that one of the sets had significantly higher variety than the other. After calculating the rank differences, 20 comparisons were selected to be rated by the human raters based on three factors:

- To find the comparisons where pairwise distance, SVS, and Nelson metrics voted differently. If all ratings agreed on the comparison, then human ratings did not add much value,
- To select the set with the highest rank difference. This was to ensure that the metric was most confident on its vote, and
- If one of the sets in the comparison had already been selected in the dataset, it was ignored. This ensured that coverage over different types of sets was obtained.

Once these 20 pairs were identified, they were given to four raters using a Qualtrics survey. In addition, two comparisons were repeated to measure the internal consistency of each raters – leading to a total of 22 comparisons. In the survey, raters were provided two sets of sketches at a time and asked to answer ``which set of milk frothers has higher variety?'', see Figure 3. This method of pairwise comparison was used because human raters are not good at giving absolute scores [40] due to differences between the internal scales of raters. The raters could choose whether Set A or Set B had higher variety or they could select the option of `cannot decide' (refer Figure 3). A percentage alignment was calculated between the expert raters to ensure independent agreement between their ratings (refer table 2). From these expert ratings, it was determined that four experts agreed on 9 out of 20 queries, while at least three experts agreed on 15 queries. Due to high agreement, these 15 queries were selected as the golden dataset.

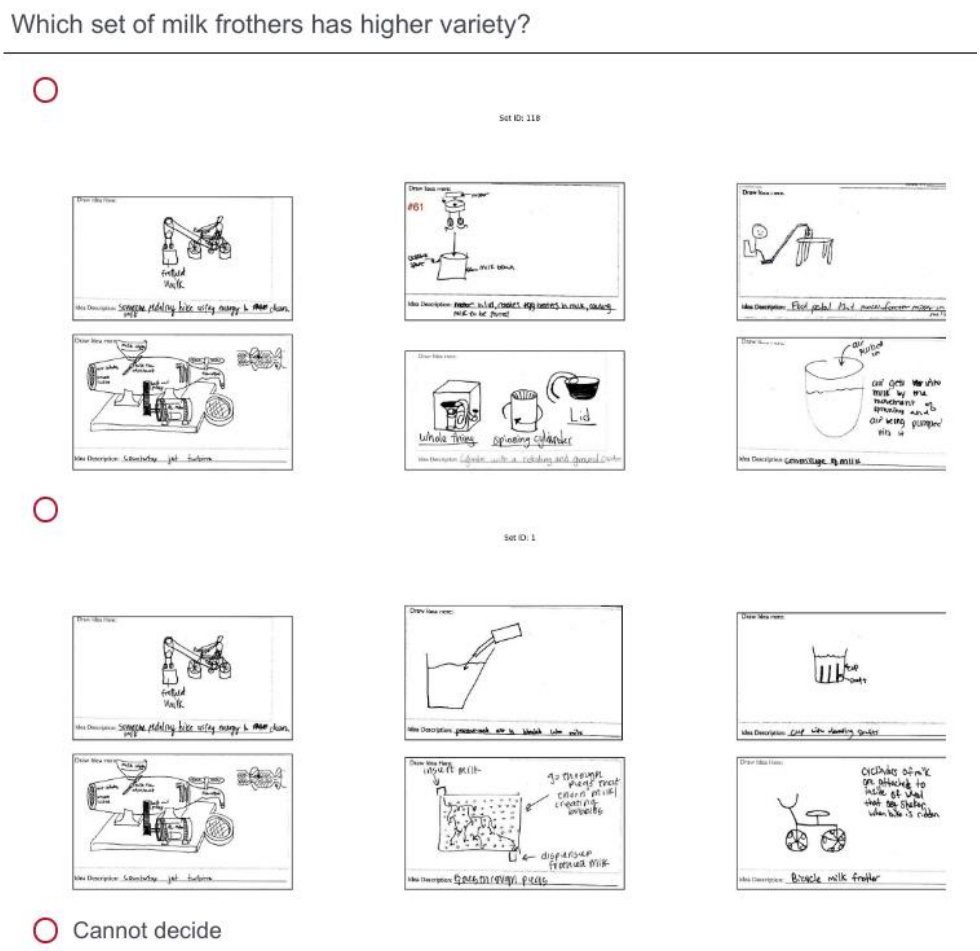

Figure 3: Screenshot of survey provided to experts to judge ideas on variety

	Rater 2	Rater 3	Rater 4
Rater 1	70%	50%	70%
Rater 2		75%	70%
Rater 3			70%

Table 2: Percentage alignment between expert raters on 20 comparisons.

Step 4: Evaluate metric validity

The golden dataset obtained in the previous step was used to test the validity of the Nelson metric. The metric's ratings were compared against the variety ratings provided by human expert raters. In order to do this, the variety for each of the 20 sets was computed using the Nelson Metric and the percentage agreement of the variety scores was compared to the survey results obtained from human expert raters. The evaluation threshold was set to 61% or 0.61 as this level of agreement is considered substantial as per the strength of agreement against Kappa Statistic proposed by Landis and Koch (Refer Table 1).

Once both metrics were computed, the ratings computed were compared using Nelson [14] variety metrics. The results showed that the variety scores obtained through the Nelson metric [14] aligned with only **33.3%** of the ground truth dataset. This is **substantially lower than the 61%** threshold.

Step 5: Refining the metric

The agreement between the metrics and the human-expert raters was found to be 33.3% as explained in the previous step. Referring back to table 1, the 33.3% agreement corresponds to a fair agreement. Due to the lack of considerable agreement between the metrics and the golden dataset, there was an attempt to refine the metrics to see if it could improve the accuracy of the metrics. Based on the results obtained, a decision was made to refine the metric using two strategies:

1. Permuting the hierarchical levels in the genealogy tree
2. Modifying the weights assigned to each level in the Nelson metric

The remainder of this section highlights how these refinements impacted the validity of these metrics.

i. Effects on variety score based on modifying the genealogical trees

As there is limited guidance on how to define each level of the genealogy tree (physical, working, embodiment, and detail), the first variety metric refinement focused on identifying the impact of modifying the tree construction by altering the definitions of each of the tree's levels (physical principle, working principle, and embodiment). Specifically, design variety scores were developed by permuting all combinations of physical principle, working principle and embodiment at different functional levels of the tree, see example Table 3.

Sl. No.	Physical Principle	Working principle	Embodiment	Variety Score (Nelson)
1	**Energy**	*Method*	Appearance	7.5
2	**Energy**	Appearance	*Method*	6.75
3	*Method*	**Energy**	Appearance	8.75
4	*Method*	Appearance	**Energy**	8
5	Appearance	Energy	*Method*	5.5
6	Appearance	*Method*	**Energy**	5.5

Table 3: The six different permutations of the genealogy tree assumed (refer Figure 2 for tree) to modify the calculation of the variety metric through the Nelson [14] metric.

ii. Effect on variety score while modifying the weights used

In addition to identifying how tree permutation impacted construct validity, the weights of the Nelson metric [108] were modified. To systematically refine the weights, the weights of the Nelson metrics were refined for trees with only three levels - Physical principle (w1), Working principle (w2) and Embodiment (w3). As explained earlier, the detail level was not included in the analysis. For this purpose, the weights were normalized between 0 to 1, as the variety scores should be scale invariant when using multiplicative weights. This implies the Nelson weights of [10, 5, and 2] are normalized to [0.59, 0.29, and 0.12]. As the sum of normalized weights equal one, the third weight for embodiment can be derived from the first two (w3= 1-w1-w2). Hence, only two-dimensional vectors of weights are reported [w1, w2].

In these experiments, the variety score for each of the 20 idea sets in the milk frother dataset were calculated for any given weight combination [w1, w2] using two methods. Using these scores, all participants were rank ordered based of the computed variety scores. Any other weights which produced the same ranking of participants as the recommended score in Nelson (weight 10, 5, 2)

implied that those weights were equivalent to the recommended weights. To measure this, the weights were perturbed and the rankings for each perturbation were calculated. Next, the correlation between default ranking and perturbed weight rankings was computed using Kendall Tau correlation.

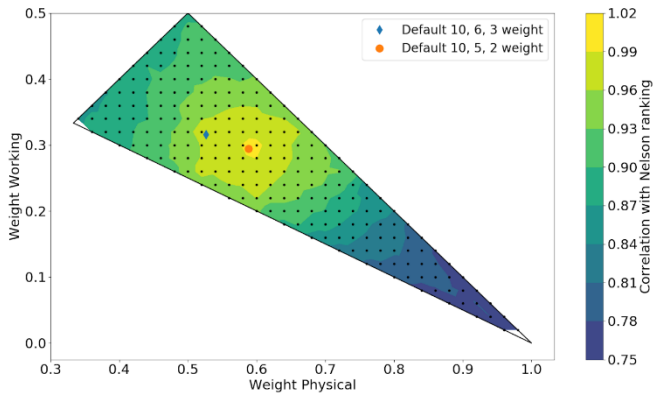

Figure 4: Kendall Tau rank correlation with default Nelson metric weights [10, 5, and 2] on using different weight combinations. For small perturbations (e.g., [10, 6 and 3], the ranking stays almost identical with rank correlation greater than 0.9

As the weights varied inside a 2-D triangle, 233 different weight combinations (as shown by black dots in Figure 4) were uniformly sampled. The figure shows the rank sensitivity for the Nelson metric [14]. One may notice from Figure 4 that slightly perturbing the weights did not change the rank ordering by variety score. For example, ideas ranked using Nelson got similar ranks for weights like [10, 8.2, and 3.4] or [10, 4.5 and 2.8] instead of [10, 5, and 2]. It was found that for small perturbations (e.g., [10, 5, and 2], the ranking stayed almost identical. Even for large changes (e.g., 10, 1, 1), the rank correlation was always greater than 0.75. This implies that for the Nelson metric, one may use any weights and the resultant rank ordering of participants does not change significantly for this dataset. Hence, the correlation between the Nelson metric and the

human expert ratings does not vary significantly from the 33.3% observed in step 3 of the framework.

However, the reader would be cautioned about carefully considering the assumptions in this study. First, the correlations were reported only for one dataset. It is possible that the metrics may have high sensitivity for some new domains. Secondly, after changing the weights, the resultant ranking was compared each time to the default Nelson weights. This shows the correlation with existing metrics. Thirdly, by using Kendall Tau correlation, the performance of the metric was measured on the basis of its ability to rank order participants. Finally, given that the detail level in the hierarchical tree was missing in ideas generated by most participants, only three levels of the tree were considered (physical principle, working principle and embodiment). Based on the results from this case study, the following key takeaways were observed:

1. In the idea sets used, trends were noticed where higher levels of abstraction were not rewarded by the genealogical tree variety metric irrespective of the type of problem.
2. There is insufficient information and examples provided by the genealogical tree variety metric and researchers who use them.
3. In the Nelson metric, the rank ordering varies the most when all weights are closer to each other.

Results and Key Takeaways from Case Study

The case study presented in this chapter highlights how the CAMbER framework can be used to identify the validity of a creativity metric. Specifically, this case study identified that there was a mismatch between the variety metric proposed by Nelson [14] and that of the human raters – the ground truth. In other words, this case study brings to question the validity of this metric if the human raters are in fact the ground truth. In addition, attempts to improve validity of the metric

through permutations of the hierarchical structure of the genealogy tree used in the Nelson method, or through modifications of the weighting structure of the computation failed. These results bring several interesting questions.

The first question that results from these findings is *why* does a mismatch exist between the Nelson and human rater methods? Based on the results obtained in this study, there is a stark mismatch between the ratings calculated by the genealogical tree method and those obtained by human-raters. Firstly, the reason for this observation could be that due to the lack of clear definitions of what constitutes to each level of abstraction for a given dataset, there might have been a misunderstanding in defining these levels on the user's end based on what constitutes to a physical principle, working principle or embodiment, resulting in the mismatch. Hence, there is a possibility that the definitions assumed were not as distinct from one another as one would hope, while evaluating a metric claiming to objectively discern an attribute of creativity (like variety). Secondly, the Kendal-Tau correlation was used as a measure to evaluate how one would rank order participants and not providing absolute variety scores as the metric actually does. The limitations of this study include using a single domain to test the sensitivity of the metrics and utilizing a subset of the large dataset of ideas (20 sets of 6 ideas each) to obtain expert ratings. 15 queries were selected to be the golden dataset based on the agreement between the experts. By using a different problem, different dataset for the same problem or different raters, the golden dataset could have varied drastically. Also, as humans are not known to provide absolute scores, pairwise comparisons were used. This is very different from how the metric computes variety by providing absolute scores for each set of data.

The second question that arises from this finding is does this case study prove that the variety metric is not valid? The answer here is – not necessarily. A single dataset based on one problem is insufficient to determine the validity of a given metric. The idea of a standardized metric

is to evaluate large datasets, both in quantity and type. Hence, it is not scientific to dismiss a metric based on the test results obtained by testing a limited dataset on one problem. There exists a possibility that the metric may have higher sensitivity in other problem types. Also, the dataset comprised of ideas generated by a single demographic population that comprised entirely of undergraduate students. The evaluation can gain more traction or be proven otherwise by evaluating ideas generated by a more diversified population of individuals including professionals, designers and students over multiple problem types in different stages of the design process.

While these results provide some insights into the validity of the variety metric and the use of the CAMbER framework, they also identify that more work is needed to ascertain the effectiveness of a given metric. The framework was utilized to test only one creativity metric. There are several aspects of creativity and metrics to measure them. In order for the methodology to be more generalizable, the CAMbER framework must be employed on other creativity metrics like novelty, quality, elegance, etc. Other directions as answered in the discussion above, are to test the metric over larger samples of data obtained from different demographics of participants, experts from varying fields, and problems from different domains such as art, psychology, design and engineering.

Sternberg [115] speaks about how each creativity measurement tap into distinct constructs of creativity and care must be taken in order to use them carefully. Recent research published by Barbot *et al.* [116] (2019) brought up an important issue to actually understand which creativity construct we are measuring. Their research reiterates the multifaceted nature of creativity and the ease with which one can wrongly generalize or use constructs of creativity interchangeably. There is no single definition that generalizes creativity. Extensive care needs to be taken that the context of operationalization of the creativity construct is always considered while measuring and comparing the particular aspect of creativity.

CHAPTER 4

QUALITATIVE STUDY ON HUMAN RATINGS: DESIGN NOVELTY

While the previous chapter highlighted the use of the CAMbER framework for validation the design variety metric, the current chapter shifts our focus to identifying *why* differences may exist between human raters and more traditional engineering design metrics, and how we might use insights in these differences to modify creativity metric computations. Specifically, this chapter highlights differences in engineering *design novelty* computations. Although there exists a plethora of metrics for measuring design creativity (see for example [12-15, 117]), these methods have been heavily criticized for their lack of generalizability across domains [20], the subjectivity of the measurements [21], the vagueness of the measurement methods [22], and the timeliness of the method for evaluating numerous concepts [118]. There is also a lack of consistency across the literature and across disciplines for which creativity metric to use and when. Recent research carried out in the domain of creativity talks about the issue how varying creativity measures cannot be used interchangeably as they point towards different aspects of creativity [115, 119, 120].

Two such disciplines that have adopted vastly different approaches for measuring creativity are the social science and engineering communities. In the social sciences, the most commonly used, albeit imperfect, method for measuring creativity has been through subjective measures. Specifically, the consensual assessment technique (CAT), put forth by Amabile [35, 69, 70] has become the ‘gold standard’ in social science research, and relies on the simple idea that an artifact

is creative only to the extent to which 'experts' in the area agree, independently, that it is creative. In contrast to social science, the majority of creativity research in engineering has focused on quantifiable measures of a concept's creativity. These metrics typically rely on breaking down design concepts into their components and then quantifying the creativity of each of these components by different means. The 'gold standard' metrics in this area were developed by Shah, Vargas-Hernandez, and Smith (SVS) [121].

Thus, the goal of the current study was to compare and contrast these two gold standard approaches by studying the creativity measurement of over 900 design ideas generated by engineering design students for a single design task and identify potential causal factors of any discrepancies. The results from this study can be used to inform both the design communities and social science communities on the utility of these approaches for measuring creativity and provide insights on whether these two disciplinary approaches are measuring the same construct of creativity and if not, how we may utilize methodologies to develop more rigorous and agreeable assessment methods. This is in line with what Nicolini, Mengis, and Swan (2012) [122] suggest, it is at intersection of disciplines that some of the most impactful and interesting investigations occur. In order to do this, a study was conducted with four human expert raters to understand what factors human raters were using to evaluate similarities or differences in design concepts. The remainder of this chapter discusses the participants, methodologies used, results, and key takeaways.

Participants and Experimental Procedure

Four raters were selected from a previous study conducted by Ahmed et al. [114], which asked 11 raters to evaluate 10 milk frother ideas in a survey, where they were provided with 360 triplet queries (all possible permutations of three sketches) and the participants had to decide whether Idea A was more similar to Idea B or Idea C (see details in Ahmed et al. [114]). This set

of design sketches was randomly sampled from the larger dataset described in Section 3.1. The internal consistency and cross-rater alignment were computed across all 11 participants. Based on this analysis, four raters showed high internal consistency and cross-rater alignment including one professor (Industrial Engineering), one post-doctoral scholar (Industrial Engineering), one Ph.D. student (Industrial Engineering) and one undergraduate student (Psychology).

These four raters were then asked to complete a second phase of ratings where each participant was provided with the same 10 idea sketches utilized in the triplet survey, printed on 8.5" x 5.5" sheets of paper, see Figure 5. The order of the ideas was randomized for each participant. The raters were required to pin the sketches on a 65" x 55" canvas (Figure 6), such that the distance between any two sketches would be proportional to how similar they were to each other. The sketches were allowed to overlap and the participants were allowed to move the sketches multiple times, until they were satisfied with the idea map created. The participants were allotted a maximum time of 30 minutes for the activity. The participants were required to think aloud and the speeches were recorded using video and audio equipment. The audio files averaged M=16 mins 48 secs (SD= 3mins 40 secs) between the four participants.

The audio was transcribed using NVivo online transcription services [97] and errors from the automatic transcriptions were manually corrected. Figure 6 shows examples of how the "milk frother" sketches were pinned on a board by one of the subjects participating in the experiment. Importantly, previous work conducted by Ahmed et al. [114], compared the maps created through this process to the maps created from the triplet survey. The current work, however, shifts the focus of the analysis to the decision-making process of the raters involved in creating these maps. In order to do this, the audio was qualitatively analyzed sentence-by-sentence using abductive content analysis [123] in NVivo [124]. Abductive content analyses was selected because it has been found to be beneficial in cases studying data with an existing theory - in this case, the novelty-tree

developed by SVS [121]) – while also taking into account the variance of data that can be obtained by participants in similar studies [125, 126]. Thus, the analysis of this data started by considering prior literature while also being responsive to the inherent characteristics of the data. In order to do this, open coding was first performed in NVIVO and then through axial coding at intersections where the participant shifted the discussion between ideas. Similar categories were grouped with the intent to understand themes and thought processes of the participants. The categories and sub-categories were directed by the content of the think-aloud recordings and prior research conducted on the same dataset of ideas [127]. The individual nodes were coded under each level of abstraction, particularly the physical principles, working principles and embodiment, as guided by the genealogical tree method proposed by SVS [121] (codes and descriptions provided in Appendices A and B). Comparing the genealogical trees used in the SVS [121] novelty metric, an analogy was assumed as to what constitutes to the physical principle, working principle and embodiment. As most of the sketches failed to dig deep enough into nitty-gritty, the detail level was ignored as suggested by the metric [121]. Two coders independently coded the data and achieved relevant inter-rater agreement to be considered for the analysis. The two raters had a high inter-rater agreement in the analysis process (Cohen's Kappa = 0.88) according to Landis's classification of Kappa [96].

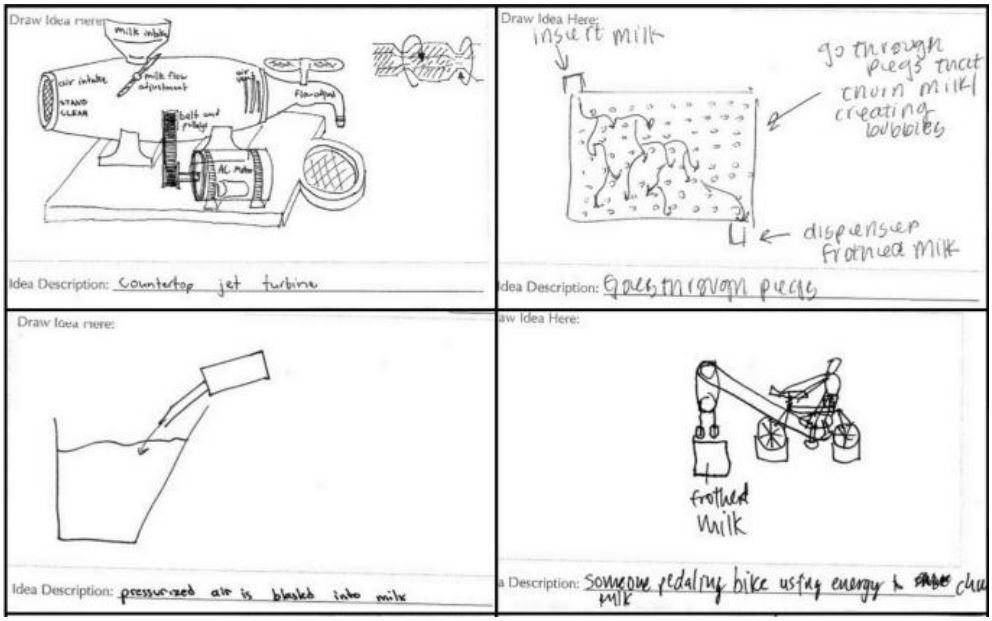

Figure 5: Example of idea sketches provided to participants. Idea 0, 1, 2 and 3 respectively.

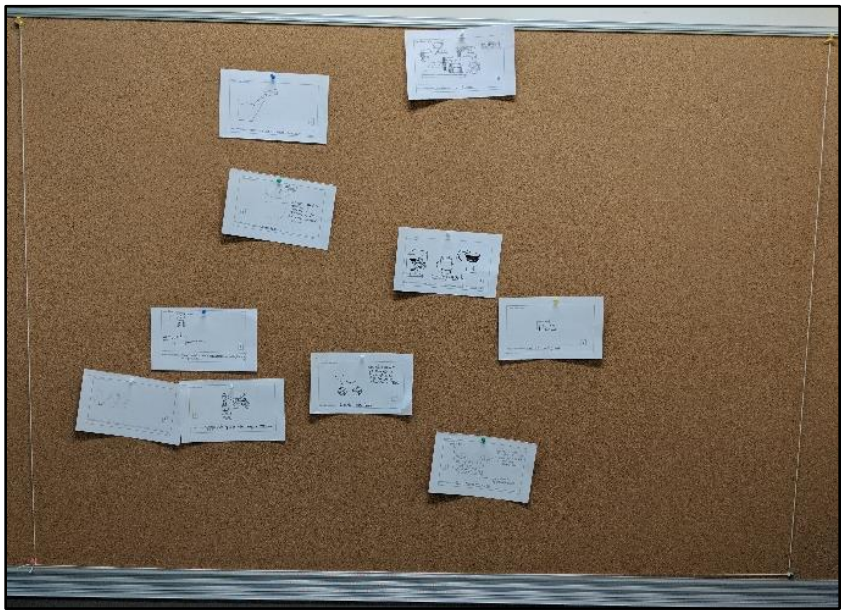

Figure 6: Ideas pinned on the board based on how "different" they are from one-another

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

In order to begin to understand ways in which we may be able to approve, the thesis turned its focus to the data gathered from the qualitative study on human ratings and the subsequent content analysis. Specifically, the content analysis identified three main discussion topics and 19 sub-topics , see Figure 7 for the cumulative normalized time raters spent discussing these topics. The remainder of this section highlights these topics and their frequencies.

As can be seen in Figure 7, the factor that was most frequently discussed was the working principle of the design (f = 90), which related to the method of frothing. This factor included discussions on air (f = 22), spinning (f = 21), movement (f = 19), vibration (f = 10), rotation (f = 6), agitation (f = 5), stirring (f = 5) and the use of a turbine (f = 2). For example, Rater 1 said "Number 5 needs to be close to 2 because it's pressurized air." Along the same lines, Rater 2 said "... idea 8 [has] a similar motion to idea 4 but it's farther away from that bicycle motion." The second most frequently discussed factor was the embodiment of the design (f = 76), which related to the physical appearance of the idea. This factor included containers (f = 23), bicycles (f = 16), beaters (f = 11), pedals (f = 11), shafts (f = 6), centrifuges (f = 4), mixers (f = 3) and pegs (f = 2). For example, Rater 4 said "So right off the bat to me ideas 3, 4, and 7 seem very similar just because they have some pedaling a bike or using a foot pedal in order to get the whole system started." Similarly, Rater 2 said "…Idea number 7. It's pretty similar to idea number 3 because they both have these pedals and they connect to a frother." Finally, the least frequently mentioned factor was the physical principle of the designs (f = 34), which related to the type of power used in the ideas. This factor included human-powered (f = 17), electricity (f = 12) and electrical-power (f = 5). For example, Rater 4 said "I'm going to move idea number 3 closer to [the other] human-power-sourced ideas."

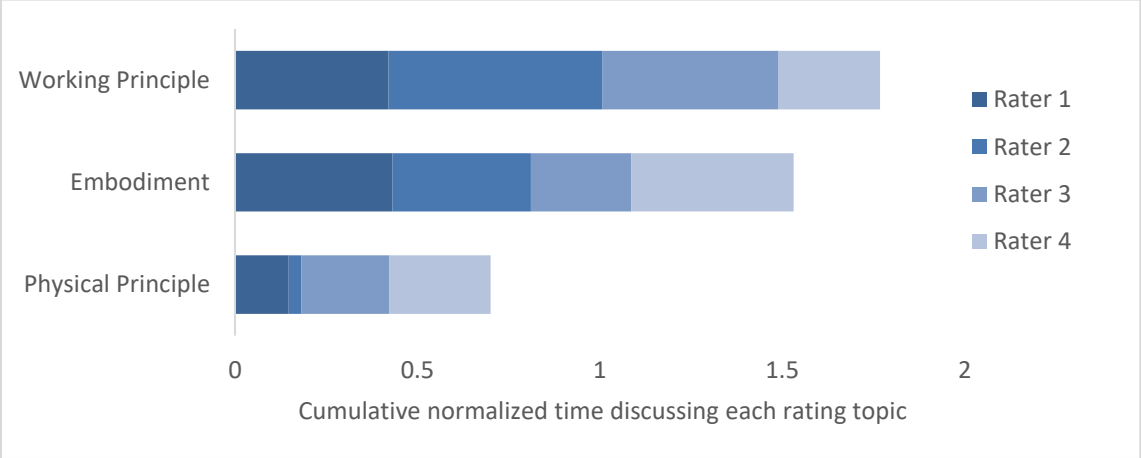

Figure 7 Cumulative normalized time spent discussing each rating topic. The normalized time is depicted in order to account for individuals who spoke more or less during the rating process.

In order to understand to identify the relative importance of these factors between the SVS and human ratings methods, the human raters cumulative normalized time spent discussing each topic was scaled to a 10-point scale and compared these weights to those assigned by the SVS method (which is out of a 10-point scale). As Table 4 demonstrates, there were large discrepancies in the relative weights assigned to these discussion topics between the two methods. In other words, the qualitative study revealed that humans are using different criteria to judge the similarity of design ideas, which may contribute to the inconsistencies observed in the novelty scores being assigned by these methods.

SVS levels	Themes from content analysis	SVS weights	Weights from content analysis
Physical Principle	Power source used	10	0.96
Working Principle	Method of frothing	6	10
Embodiment	Form	3	8.66

Table 4: Weights assigned by the SVS method and from the content analysis. The content analysis weights were determined by taking the cumulative normalized time the raters spent discussing each topic and transposing that to a 10-point scale.

Results: Can human ratings be used to improve the validity of the design novelty metric?

In order to answer this question, the audio transcriptions were analyzed to discern the amount of time spent by each expert rater talking about each topic as they mapped their ideas based on similarity. The cumulative normalized time spent by each rater on the topics: power source used, method of frothing and form were obtained. The time discussing the topics was transposed on a 10-point scale (as SVS novelty scale has a maximum weight of 10) while maintaining the ratio between the levels of abstraction as obtained through the content analysis. In essence, based on the weights obtained, the novelty ratings were calculated as prescribed by SVS [16] using these new weights as depicted in table 4. The original SVS novelty metric and the new SVS novelty metric (created by replacing the existing weights of 10, 6 and 1 with the new weights of 0.96, 10 and 8.66 were compared against CAT ratings obtained by expert assessment.

There was a significant but weak positive correlation between the novelty score of an idea rated by SVS score and CAT score, $r = .10$, $p < .01$, such that higher an idea scored on the SVS novelty metric, the more likely it is that it also scored high on the CAT score. By contrast when

using the SVS metric with new weights from the content analysis, this correlation disappears, resulting in a non-significant, near-zero correlation between new SVS score and CAT score, r = .03, p = .31.

Due to the low correlation between either of the SVS ratings (original with weights of 10, 6 and 1, and the new SVS metric with weights of .96, 10 and 8.66 from the content analysis) and CAT, the results obtained from both the SVS ratings were compared against each other. There was a significant strong positive correlation between the novelty ratings of ideas using the old and new SVS metric, r = .89, p < .001, such that the higher an idea scored for novelty on the old SVS metric, the higher its rating would be on the new SVS metric.

Discussion

This chapter attempted to review the possibility of using human expert raters to validate metrics that have been widely utilized to compute design novelty. This part of the thesis gives us a possible method to refine existing metrics based on human raters and how they interpret similarities and differences in design concepts. In spite of using human raters to calibrate existing genealogical tree metric by modifying the weights, little to no improvement was observed with respect to the correlation of the ratings with the gold standard (CAT) ratings. In fact, the correlation and significance of correlation dropped between SVS and CAT on modifying the weights. These results are similar to what was observed in the framework evaluation of variety metrics. It appears that we are using the genealogical tree metrics without substantially evaluating the validity, reasoning behind the usage of tree frameworks, and the arbitrary weights suggested.

It is to be noted that the experiment was conducted on a single dataset of ideas on a single type of problem. This methodology needs to be tested with ideas generated by a population with diverse backgrounds over multiple problems to check if these findings can be repeated. Due to the

time constraints and physical space limitations, only 10 ideas were chosen to be mapped on the board. Adding a larger set of ideas maybe helpful in better capturing the thought process behind experts making decisions on similarity of ideas. Another important point to be noted is that, when human raters are required to compare multiple ideas at a time, with each idea added, there is a higher cognitive load on the expert to remember the relationship between two ideas while also considering how moving one idea changes the relationship between all ideas. Having experts compare fewer ideas at a time over multiple sets of ideas could result in higher richness of the data obtained. These are a few reasons why the study might have resulted in a mismatch between the CAT and SVS ratings in spite of modifying the weights. Using these readily available new weights produced by this study on other diverse and larger datasets could help determine the effectiveness of this modification of the SVS metric. Use of physical products rather than sketches of ideas might reduce the extent of abstraction on the experts' end, and in turn modify the weights to produce more appropriate ratings. The results obtained in this study warrant a need to extensively evaluate the metrics utilized across engineering design research today.

CHAPTER 5

CONCLUSION

While creativity metrics have been developed and used over the past three decades, there exists a lack of consistency in the approach used to quantify different aspects of creativity. When metrics claiming to be measuring the same construct of creativity produce different results, it is imperative for researchers to establish a universal ground truth to compare and quantify metrics against. The goals of this thesis was to create and execute a framework to test the accuracy of creativity metrics currently used in engineering design. The framework establishes a method through a stepwise process to identify a construct for review and determine the type of validation. Then, using the variety metric as an example, the framework walks the researcher through establishing the ground truth based on evaluations from expert raters who independently agree based on subjective ratings and evaluating metric for validity. The framework also provides opportunity for metric refinement if the results from the metric modification are unsatisfactory.

The second goal of the thesis was to compare two gold standards of creativity metrics utilized in the social (CAT) and engineering sciences (SVS) and to demonstrate how we could leverage the advantages from both these schools of thought to develop more rigorous and agreeable metrics. The Consensual Assessment Technique is expert driven, while the genealogical tree metric is computationally driven. The success of experiments of this nature can give birth to metrics measuring attributes of creativity that can be quickly calculated, while also possessing the strength of domain-specific expert knowledge which is known to capture the abstract nature of creativity effectively. Importantly, the results from this study encourage us to regularly check the accuracy

and precision of metrics used in engineering design studies. We observe regular improvements to existing metrics on their theoretical implementations, however, there exists a disconnect between what the metrics are measuring and the ground truth when it comes to reporting new or existing metrics. The study aimed at producing a possible method to refine existing metrics based on how human raters interpret similarities and differences in design concepts focusing on originality or novelty of ideas. It was found that the modified weights, though drastically modified, did not change the metric's correlation with the ratings obtained through subjective assessment of creativity. This brings to light a very important question to the researchers, as to why we are utilizing the weights if major changes to them does not result in significant changes in the ratings.

While thinking about refining scales, there have also been examples of moving to new ways of measuring the same attributes. The following example speaks about utilizing existing ground truths to create a new reference for measurement. The International Prototype Kilogram has been said to have lost 50 micrograms in mass [61]. As such, on May 20, 2019, it was decided that the international standards will be moving away from the IPK, as the prototype has been losing weight in spite of maintaining it undisturbed, in the same location, encased in trio vacuum-sealed bell jars. The current plan proposed by researchers is to move to universal fundamental constants like Planck's constant using a Kibble balance (also referred to as Planck balance/Watt balance [128]) or Avogadro's number using X-ray crystal density to measure mass [61]. The new definition of Kilogram is as follows:

"The kilogram is defined by taking the fixed numerical value of the Planck constant h to be 6.62607015×10−34 when expressed in the unit J s, which is equal to kg m2 s−1, where the meter and the second are defined in terms of c and ΔνCs" [61, 129] (where c= the speed of light in vacuum; 299,792,458 meters per second; Δν = the unperturbed ground state hyperfine transition frequency of the caesium-133 atom; ΔνCs is 9,192,631,770 hertz) [129]. By changing the way the

kilogram is measured, researchers have refined the metric to ensure the highest possible accuracy with the technology available today [61].

For decades we've inherently assumed that the techniques used to measure creativity are accurate. Due to discrepancies found between metrics over the past 5-10 years, we are forced to test methods for their validity, and calibrate them in order to ensure the best ideas can be harvested to create better products and services. This research work is a stepping stone towards establishing new methodologies to validate creativity while providing room for refinement. As Isaac Newton once said "*If I had seen further than others, it is by standing upon the shoulders of giants*", this work does not aim at disproving existing work, but at improving the methods of measurement through validation and refinement. This study is an example as to how multiple schools of thought can be leveraged to extract the sought-out qualities from different metrics to create a more rigorous tool to quantify the abstract construct of creativity. There is potential for this methodology to be utilized to assess other measurement devices beyond metrics that are used to appraise creativity.

Limitations and Future Work

This is a first step towards questioning the validity of metrics widely used in the field of engineering and design to quantify creativity. This thesis proposes two different approaches to utilize human expertise in calibrating objective metrics used in engineering and design to quantify creativity. The study however, comes with its set of limitations as discussed earlier. The datasets used in the experiments are from a single problem type and specific demographic of undergraduate student population. In both studies only one metric was tested using each methodology. The definitions used to signify each level of abstraction were assumed based on the users' interpretation of the metric as the variety and novelty metric fail to mention what constitutes to physical principle, working principle and embodiment for varied problem types.

Next, the ability to discern who an expert is, is tricky when it comes to a diversified field like design or engineering. While the problems explored here were relatively simple, the results are likely to be exasperated in more complex problems like those found in engineering design and systems engineering [130, 131]. Given the importance of expertise in the rating process [35, 69, 132] and the findings of the study that clearly identify difference between expert and quasi-expert raters in engineering design quality and creativity ratings, it is important to explore training methods for improving the viability and utility of rating assessments. This is particularly important in engineering due to the use of novices or quasi-experts in published articles (see for example [133]), the difficulty in quantifying expertise in engineering domains which are multi-disciplinary in nature, and the time required by experts to perform these assessments (which often makes expert ratings unattainable).

Future work should aim to include a larger, more diversified dataset from a variety of problems. There is a need to concentrate on methods to ease the ability to arrive at the ground truth to ensure boundless adaptation of the framework across engineering and social psychology. And in order to successfully do that, the definitions of each level of abstraction have to be clearly distinguished especially while trying to discern abstract concepts such as creativity. The validity of the methodology rests on our ability to find experts and rely on their skills to produce ground truth results. A process to determine domain relevant expertise is pivotal to the success of this framework.

With the amount of progress made in the field of artificial intelligence and machine learning, there is potential in looking into AI agents that can learn the parameters humans use to evaluate creativity and replicate creativity ratings. Genealogical tree metrics are used as objective metrics in the field of engineering design. CAT uses domain-expertise to analyze creativity. If we can train AI agents to mimic domain-specific-human-expertise by observing large amounts of

creativity ratings, we could arrive at a hybrid metric that could potentially combine the advantages of subjective and objective creativity metrics.

In a special issue on creativity assessment recently published (2019) by Barbot *et al.* [116] an important notion is emphasized on how there is a high extent of variability in assessing creativity. There is a need to identify an optimal standard measurement procedure for each construct of creativity in order to ensure homogeneity [134]. These efforts are necessary in order to attain repeatable and reliable creativity studies in psychology and engineering. A testing tool like CAMbER can aid in identifying ground truths for each individual construct of creativity and verify if a given creativity assessment tool is effective in measuring a particular construct of creativity and to what extent.

These experiments in spite of their results are directions towards testing the validity of metrics without blindly adopting them. Albert Einstein said: "*The smartest people on the planet are often the ones who ask the most questions*". By not blindly following methods laid out by past research, this work provides a ray of hope in questioning the way we measure creativity today. Future work in this direction will add value to these methodologies, and help in comprehensively testing creativity metrics.

APPENDIX A – CODEBOOK FOR CONTENT ANALYSIS

Levels of Abstraction	**Description**	**Example**
Physical Principle	The participant's response indicating type of power source that was used to power the product	"All right, I am sticking with the idea here that these are manually powered. Nine, three and seven are manually powered."
Working Principle	The participant's response indicating the type of motion used by the product	"This one is spinning so it's close to that one, idea number eight. It's a spinning cylinder."
Embodiment	The participant's response indicating what the product looked like	"…and they use a bicycle so I am going to use these as my starting points."

APPENDIX B – CODE DESCRIPTION FOR CONTENT ANALYSIS

Levels of Abstraction	Themes	Description
Physical Principle	Power	The participant discusses differentiation with reference to power/source of power utilized. (e.g. power, powered, source, sources)
	Human-powered	The participant discusses ideas with reference to humans used to power the frothing of milk. (e.g. human, foot, manually, manual)
	Electricity	The participant discusses ideas with reference to electricity used to power the frothing of milk. (e.g. electrically, electricity, motor)
Working Principle	Agitation	The participant discusses ideas with reference to agitation used as the mechanism to froth milk. (e.g. agitate, agitated, agitating, agitation, agitator)
	Air	The participant discusses ideas with reference to air pressure used as the mechanism to froth milk. (e.g. air, pressure, pressurized, blasted)
	Move	The participant discusses ideas with reference to movement of objects/milk as a mechanism to froth milk. (e.g. Move, movement, moving, motion, push, pushed, pushing, shake, shaking)

Working Principle (Continued)	Rotate	The participant discusses ideas with reference to rotation of objects/milk to froth milk. (e.g. Rotate, rotates, rotating)
	Spinning	The participant discusses ideas with reference to spinning of objects/milk as a mechanism to froth milk. (e.g. Spin, spinning, spins)
	Stirring	The participant discusses ideas with reference to stirring of objects/milk as a mechanism to froth milk. (e.g. Stir, stirring, stirred)
	Turbine	The participant discusses ideas with references to a jet/turbine used to froth milk. (e.g. Jet, turbine)
	Vibration	The participant discusses ideas with reference to vibration used as a mechanism to froth milk. (e.g. Vibrate, vibrating, vibration)
Embodiment	Beater	The participant discusses ideas that involve beaters.
	Bicycle	The participant discusses ideas that involve a bicycle. (e.g. Bicycle, bike)
	Centrifuge	The participant discusses ideas that involve centrifuges.
	Containers	The participant discusses ideas that involve containers (e.g. containers, cups and bowls)
	Pedal	The participant discusses ideas that involve pedals. (e.g. pedal, pedals, pedaling)

Embodiment (Continued)	Mixer	The participant discusses ideas with reference to a mixer used as a mechanism to froth milk.
	Pegs	The participant discusses ideas that involve pegs.
	Shaft	The participant discusses ideas that involve shafts.

BIBLIOGRAPHY

1. Amabile, T.M., *Creativity in context: Update to the social psychology of creativity*. 1996: Hachette UK.
2. Brands, R.F. and M.J. Kleinman, *Robert's Rules of Innovation: A 10-Step Program for Corporate Survival*. 2010: John Wiley & Sons.
3. Kulkarni, S., et al., *Evaluation of collaborative (C-Sketch) as an idea generation technique for engineering design.* Journal of Creative Behavior, 2000. **35**(3): p. 168-198.
4. Toh, C.A. and S.R. Miller. *Exploring the utility of product dissection for early-phase idea generation*. in *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. 2013. American Society of Mechanical Engineers.
5. Daly, S.R., et al., *Design heuristics in engineering concept generation.* Journal of Engineering Education, 2012. **101**(4): p. 601-629.
6. Daly, S.R., et al., *Assessing design heuristics for idea generation in an introductory engineering course.* International Journal of Engineering Education, 2012. **28**(2): p. 463.
7. VanGrundy, A., *Techniques of Structured Problem Solving*. 2nd ed. 1988, New York, NY: Van Nostrand Reinhold Company.
8. Altshuller, G.S., *Creativity as an exact science: the theory of the solution of inventive problems*. 1984: Gordon and Breach.
9. Eberle, B., *Scamper on: Games for imagination development*. 1996: Prufrock Press Inc.
10. Osborn, A., *Applied Imagination*. 1957, New York, NY: Scribner.
11. Jablokow, K., et al., *Ideation Variety in Mechanical Design: Examining the Effects of Cognitive Style and Design Heuristics.* 2015(57106).
12. Hernandez, N., G. Okudan Kremer, and L.C. Schmidt, *Effectiveness metrics for ideation: Merging genealogy trees and improving novelty metric*, in *International Design Engineering Technical Conferences*. 2012: Chicago, IL. p. 85-93.
13. Peeters, J., et al., *Refined metrics for measuring novelty in ideation.* Proc. IDMME Virtual Concept 2010, 2010.
14. Nelson, B. and J. Yen, *Refined metrics for measuring ideation effectiveness.* Design Studies, 2009. **30**(6): p. 737-743.
15. Johnson, T.A., et al. *Comparison and Extension of Novelty Metrics for Problem-Solving Tasks*. in *ASME 2016 International Design Engineering Technical Conferences & Computers and Information Engineering Conference*. 2016. Charlotte, NC: ASME.
16. Shah, J.J., S.M. Smith, and N. Vargas-Hernandez, *Metrics for measuring ideation effectiveness.* Design studies, 2003. **24**(2): p. 111-134.
17. Chakrabarti, A. and P. Khadilkar. *A measure for assessing product novelty*. in *DS 31: Proceedings of ICED 03, the 14th International Conference on Engineering Design, Stockholm*. 2003.
18. Lopez-Mesa, B. and R. Vidal. *Novelty metrics in engineering design experiments*. in *DS 36: Proceedings DESIGN 2006, the 9th International Design Conference, Dubrovnik, Croatia*. 2006.
19. Sarkar, P. and A. Chakrabarti, *Assessing design creativity.* Design Studies, 2011. **32**(4): p. 348-383.

20. Baer, J., *Domain specificity and the limits of creativity theory.* The Journal of Creative Behavior, 2012. **46**(1): p. 16-29.
21. Casakin, H. and S. Kreitler, *The nature of creativity in design.* Studying Designers, 2005. **5**: p. 87-100.
22. Williams, A.P., M.J. Ostwald, and H.H. Askland, *The Relationship between Creativity and Design and Its Implication for Design Education.* Design Principles & Practice: An International Journal, 2011. **5**(1).
23. DeMarco, T., *Controlling software projects: Management, measurement, and estimates*. 1986: Prentice Hall PTR Upper Saddle River, NJ, USA.
24. Barrett, G.V., *Research models of the future for industrial and organizational psychology.* Personnel Psychology, 1972.
25. Cronbach, L.J. and P.E. Meehl, *Construct validity in psychological tests.* Psychological bulletin, 1955. **52**(4): p. 281.
26. Bagozzi, R.P., Y. Yi, and L.W. Phillips, *Assessing construct validity in organizational research.* Administrative science quarterly, 1991: p. 421-458.
27. Mitchell, T.R., *An evaluation of the validity of correlational research conducted in organizations.* Academy of Management review, 1985. **10**(2): p. 192-205.
28. Turner, S.M., et al., *APA's guidelines for test user qualifications: an executive summary.* American Psychologist, 2001. **56**(12): p. 1099.
29. Cascio, W.F., *Applied psychology in human resource management.* 1998.
30. Frey, B.B., *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. 2018: SAGE Publications.
31. Conway, J.M., R.A. Jako, and D.F. Goodman, *A meta-analysis of interrater and internal consistency reliability of selection interviews.* Journal of applied psychology, 1995. **80**(5): p. 565.
32. Beck, A.T. and R.A. Steer, *Internal consistencies of the original and revised Beck Depression Inventory.* Journal of clinical psychology, 1984. **40**(6): p. 1365-1367.
33. Streiner, D.L., *Starting at the beginning: an introduction to coefficient alpha and internal consistency.* Journal of personality assessment, 2003. **80**(1): p. 99-103.
34. Messick, S., *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning.* American psychologist, 1995. **50**(9): p. 741.
35. Amabile, T., *Social psychology of creativity: A consensusual assessment technique.* Journal of Personality and Social Psychology, 1982. **43**: p. 997-1013.
36. Torrance, E.P., *Torrance tests of creative thinking: Norms-technical manual: Verbal tests, forms a and b: Figural tests, forms a and b*. 1966: Personal Press, Incorporated.
37. Price, J.L., *Handbook of organizational measurement.* International journal of manpower, 1997. **18**(4/5/6): p. 305-558.
38. Carmines, E.G. and R.A. Zeller, *Reliability and validity assessment*. Vol. 17. 1979: Sage publications.
39. Bollen, K.A., *Structural equations with latent variables*. Vol. 210. 2014: John Wiley & Sons.
40. Kirk, J., M.L. Miller, and M.L. Miller, *Reliability and validity in qualitative research*. Vol. 1. 1986: Sage.
41. Trochim, W.M. and J.P. Donnelly, *Research methods knowledge base*. Vol. 2. 2001: Atomic Dog Publishing Cincinnati, OH.
42. Saunders, R., *Curious design agents and artificial creativity.* 2002.
43. Zheng, X. and S.R. Miller. *Linking Creativity Measurements to Product Market Favorability: A Data-Mining Approach*. in *ASME 2017 International Design Engineering*

Technical Conferences and Computers and Information in Engineering Conference. 2017. American Society of Mechanical Engineers.
44. Crosland, M., *The Congress on Definitive Metric Standards, 1798-1799: The First International Scientific Conference?* Isis, 1969. **60**(2): p. 226-231.
45. Giacomo, P., *The new definition of the meter.* American Journal of Physics, 1984. **52**(7): p. 607-613.
46. Slaney, K., *Validating psychological constructs: Historical, philosophical, and practical dimensions*. 2017: Springer.
47. Association, A.E.R., et al., *Standards for educational and psychological testing*. 1985: American Educational Research Association.
48. Kane, M., *Content-related validity evidence in test development.* Handbook of test development, 2006. **1**: p. 131-153.
49. Messick, S., *Validity In. R. Linn (Ed.) Educational measurement (13-103)*. 1989, New York: Macmillan Publishing.
50. Haynes, S.N., D. Richard, and E.S. Kubany, *Content validity in psychological assessment: A functional approach to concepts and methods.* Psychological assessment, 1995. **7**(3): p. 238.
51. Yaghmaie, F., *Content validity and its estimation.* Journal of Medical Education, 2003. **3**(1).
52. Lawshe, C.H., *A quantitative approach to content validity 1.* Personnel psychology, 1975. **28**(4): p. 563-575.
53. Berk, R.A., *Importance of expert judgment in content-related validity evidence.* Western journal of nursing research, 1990. **12**(5): p. 659-671.
54. Tenopyr, M.L., *CONTENT–CONSTRUCT CONFUSION 1.* Personnel Psychology, 1977. **30**(1): p. 47-54.
55. Price, P.C., R. Jhangiani, and I.-C.A. Chiang, *Research methods in psychology*. 2015: BCCampus.
56. Cramond, B., et al., *A report on the 40-year follow-up of the Torrance Tests of Creative Thinking: Alive and well in the new millennium.* Gifted Child Quarterly, 2005. **49**(4): p. 283-291.
57. Plucker, J.A., *Is the proof in the pudding? Reanalyses of Torrance's (1958 to present) longitudinal data.* Creativity Research Journal, 1999. **12**(2): p. 103-114.
58. Kim, K.H., *Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT).* Creativity research journal, 2006. **18**(1): p. 3-14.
59. Torrance, E.P. and T. Wu, *A comparative longitudinal study of the adult creative achievements of elementary school children identified as highly intelligent and as highly creative.* Creative Child and Adult Quarterly, 1981. **6**(2): p. 71-76.
60. Almeida, L.S., et al., *Torrance Test of Creative Thinking: The question of its construct validity.* Thinking skills and creativity, 2008. **3**(1): p. 53-58.
61. Haller, J., et al., *The redefinition of the unit „kilogram "–what does it mean for weighing technology users?* 2019.
62. Davis, R.S., P. Barat, and M. Stock, *A brief history of the unit of mass: continuity of successive definitions of the kilogram.* Metrologia, 2016. **53**(5): p. A12.
63. Mouchiroud, C. and A. Bernoussi, *An empirical study of the construct validity of social creativity.* Learning and Individual Differences, 2008. **18**(4): p. 372-380.
64. Brackstone, G., *Managing data quality in a statistical agency.* 2003.
65. Haidegger, T., et al. *The importance of accuracy measurement standards for computer-integrated interventional systems*. in *EURON GEM Sig Workshop on The Role of Experiments in Robotics Research at IEEE ICRA*. 2010.

66. Schoene, B., et al., *Precision and accuracy in geochronology.* Elements, 2013. **9**(1): p. 19-24.
67. Harvey, E.S. and M.R. Shortis, *Calibration stability of an underwater stereo-video system: implications for measurement accuracy and precision.* Marine Technology Society Journal, 1998. **32**(2): p. 3-17.
68. Kalton, G. *How important is accuracy*. in *Proceedings of Statistics Canada Symposium, Canada*. 2001. Citeseer.
69. Amabile, T., *Brilliant but cruel: perceptions of negative evaluators.* Journal of Experimental Psychology, 1983. **19**(2): p. 146-156.
70. Amabile, T., *Creativitiy in Context*. 1996, Boulder, Colorado: Westview Press.
71. Baer, J., J.C. Kaufman, and C.A. Gentile, *Extension of the consensual assessment technique to nonparallel creative products.* Creativity research journal, 2004. **16**(1): p. 113-117.
72. Kaufman, J.C., et al., *Creativity stereotypes and the consensual assessment technique.* Creativity Research Journal, 2010. **22**(2): p. 200-205.
73. Jagtap, S., *Design creativity: refined method for novelty assessment.* International Journal of Design Creativity and Innovation, 2019. **7**(1-2): p. 99-115.
74. Chulvi, V., et al., *Comparison of the degree of creativity in the design outcomes using different design methods.* Journal of Engineering Design, 2012. **23**(4): p. 241-269.
75. Benedek, M., et al., *Assessment of divergent thinking by means of the subjective top-scoring method: Effects of the number of top-ideas and time-on-task on reliability and validity.* Psychology of aesthetics, creativity, and the arts, 2013. **7**(4): p. 341.
76. Chiu, I. and F.A. Salustri, *Evaluating design project creativity in engineering design courses.* Proceedings of the Canadian Engineering Education Association (CEEA), 2010.
77. Green, M., C.C. Seepersad, and K. Hölttä-Otto. *Crowd-sourcing the evaluation of creativity in conceptual design: A pilot study*. in *ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. 2014. American Society of Mechanical Engineers.
78. Van Der Velde, F., et al. *A Semantic Map for Evaluating Creativity*. in *ICCC*. 2015.
79. Carroll, E.A., et al. *Creativity factor evaluation: towards a standardized survey metric for creativity support*. in *Proceedings of the seventh ACM conference on Creativity and cognition*. 2009. ACM.
80. O'Quin, K. and S.P. Besemer, *The development, reliability, and validity of the revised creative product semantic scale.* Creativity Research Journal, 1989. **2**(4): p. 267-278.
81. Ritchie, G. *Assessing creativity*. in *Proc. of AISB'01 Symposium*. 2001. Citeseer.
82. Hogan, T.P., A. Benjamin, and K.L. Brezinski, *Reliability methods: A note on the frequency of use of various types.* Educational and psychological measurement, 2000. **60**(4): p. 523-531.
83. Henson, R.K., *Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha.(Methods, plainly speaking).* Measurement and evaluation in counseling and development, 2001. **34**(3): p. 177-190.
84. Cropley, A.J., *Defining and measuring creativity: Are creativity tests worth using?* Roeper review, 2000. **23**(2): p. 72-79.
85. Zenasni, F., M. Besancon, and T. Lubart, *Creativity and tolerance of ambiguity: An empirical study.* The Journal of Creative Behavior, 2008. **42**(1): p. 61-73.
86. Clapham, M.M., *The convergent validity of the Torrance Tests of Creative Thinking and creativity interest inventories.* Educational and Psychological Measurement, 2004. **64**(5): p. 828-841.

87. Amabile, T.M., et al., *Assessing the work environment for creativity.* Academy of management journal, 1996. **39**(5): p. 1154-1184.
88. Hu, W. and P. Adey, *A scientific creativity test for secondary school students.* International Journal of Science Education, 2002. **24**(4): p. 389-403.
89. Gough, H.G., *California psychological inventory*. 1956.
90. Gough, H.G., *California psychological inventory: Administrator's guide*. 1987: Consulting Psychologists Press.
91. Myers, I.B., M.H. McCaulley, and R. Most, *Manual, a guide to the development and use of the Myers-Briggs type indicator*. 1985: consulting psychologists press.
92. Myers, I.B., et al., *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Vol. 3. 1998: Consulting Psychologists Press Palo Alto, CA.
93. Kirton, M., *Adaptors and innovators: A description and measure.* Journal of applied psychology, 1976. **61**(5): p. 622.
94. Kirton, M.J., *Adaptors and innovators: Cognitive style and personality*. Frontiers of creativity research, 1987: p. 282-304.
95. Fleenor, J.W. and S. Taylor, *Construct validity of three self-report measures of creativity.* Educational and Psychological Measurement, 1994. **54**(2): p. 464-470.
96. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data.* biometrics, 1977: p. 159-174.
97. QSR. *NVivo Transcription services*. NVivo 2018; Available from: https://www.qsrinternational.com/nvivo/nvivo-products/transcription.
98. Fontenot, N.A., *Effects of training in creativity and creative problem finding upon business people.* The Journal of Social Psychology, 1993. **133**(1): p. 11-22.
99. Ward, W.C., *Creativity in young children.* Child Development, 1968: p. 737-754.
100. Guilford, J.P. and R. Hoepfner, *The analysis of intelligence*. 1971: McGraw-Hill Companies.
101. Brennan, M.A., *Dance creativity measures: a reliability study.* Research Quarterly for Exercise and Sport, 1983. **54**(3): p. 293-295.
102. Jagtap, S. *Assessing design creativity: Refinements to the novelty assessment method*. in *DS 84: Proceedings of the DESIGN 2016 14th International Design Conference*. 2016.
103. Milton, M. and I. Mills, *Amount of substance and the proposed redefinition of the mole.* Metrologia, 2009. **46**(3): p. 332.
104. Mills, I.M., et al., *Redefinition of the kilogram, ampere, kelvin and mole: a proposed approach to implementing CIPM recommendation 1 (CI-2005).* Metrologia, 2006. **43**(3): p. 227.
105. Taylor, B.N., *Molar mass and related quantities in the New SI.* Metrologia, 2009. **46**(3): p. L16.
106. Hernandez, N.V., G.E. Okudan, and L.C. Schmidt. *Effectiveness metrics for ideation: Merging genealogy trees and improving novelty metric*. in *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. 2012. American Society of Mechanical Engineers.
107. Toh, C.A. and S.R. Miller, *Choosing creativity: the role of individual risk and ambiguity aversion on creative concept selection in engineering design.* Research in Engineering Design, 2016. **27**(3): p. 195-219.
108. Nelson, B.A., et al., *Refined metrics for measuring ideation effectiveness.* Design Studies, 2009. **30**(6): p. 737-743.
109. Torrance, E., *Predictive validity of the Torrance tests of creative thinking.* The Journal of Creative Behavior, 1972. **6**(4): p. 236-262.

110. Henderson, D., et al. *A Comparison of Variety Metrics in Engineering Design*. in *ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. 2017. American Society of Mechanical Engineers.
111. Starkey, E.M., et al. *Confidently Exploring the Solution Space: The Within-Subject Effects of Product Dissection on Design Variety and Creative Self-Efficacy*. in *ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. 2018. American Society of Mechanical Engineers.
112. Runco, M.A. and S. Acar, *Divergent thinking as an indicator of creative potential.* Creativity Research Journal, 2012. **24**(1): p. 66-75.
113. Plucker, J.A., M. Qian, and S.L. Schmalensee, *Is what you see what you really get? Comparison of scoring techniques in the assessment of real-world divergent thinking.* Creativity Research Journal, 2014. **26**(2): p. 135-143.
114. Ahmed, F., et al., *Interpreting Idea Maps: Pairwise comparisons reveal what makes ideas novel.* Journal of Mechanical Design, 2019. **141**(2): p. 021102.
115. Sternberg, R.J., *What's wrong with creativity testing?* The Journal of Creative Behavior, 2018.
116. Barbot, B., R.W. Hass, and R. Reiter-Palmon, *Creativity assessment in psychological research:(Re) setting the standards.* Psychology of Aesthetics, Creativity, and the Arts, 2019. **13**(2): p. 233.
117. Borgianni, Y., G. Cascini, and F. Rotini, *Assessing creativity of design projects: Criteria for the service engineering field.* International Journal of Design Creativity and Innovation, 2013. **1**(3): p. 131-159.
118. Gosnell, C.A. and S.R. Miller, *A novel method for assessing design concept creativity using single-word adjectives and semantic similarity*, in *ASME Design Engineering Technical Conferences*. 2014: Buffalo, NY.
119. Glăveanu, V.P., *Measuring creativity across cultures: Epistemological and methodological considerations.* Psychology of Aesthetics, Creativity, and the Arts, 2019. **13**(2): p. 227.
120. Hornberg, J. and R. Reiter-Palmon, *Creativity and the big five personality traits: Is the relationship dependent on the creativity measure?* 2017.
121. Shah, J., S. Smith, and N. Vargas-Hernandez, *Metrics for Measuring Ideation Effectiveness.* Design Studies, 2003. **24**: p. 111-124.
122. Nicolini, D., J. Mengis, and J. Swan, *Understanding the role of objects in cross-disciplinary collaboration.* Organization science, 2012. **23**(3): p. 612-629.
123. Timmermans, S. and I. Tavory, *Theory construction in qualitative research: From grounded theory to abductive analysis.* Sociological theory, 2012. **30**(3): p. 167-186.
124. Software, N.Q.D.A., *QSR International Pty Ltd. Version 10.* 2012.
125. Charmaz, K. and L. Belgrave, *Qualitative interviewing and grounded theory analysis.* The SAGE handbook of interview research: The complexity of the craft, 2012. **2**: p. 347-365.
126. Creswell, J.W. and D.L. Miller, *Determining validity in qualitative inquiry.* Theory into practice, 2000. **39**(3): p. 124-130.
127. Starkey, E., C.A. Toh, and S.R. Miller, *Abandoning creativity: The evolution of creative ideas in engineering design course projects.* Design Studies, 2016.
128. Chao, L., et al., *A LEGO Watt balance: An apparatus to determine a mass based on the new SI.* American Journal of Physics, 2015. **83**(11): p. 913-922.

129. Lopac, V. and D. Hrupec, *What exactly are the new definitions of kilogram and other SI units?* The Physics Teacher, 2012.
130. Kaufman, J.C. and J. Baer, *Beyond new and appropriate: Who decides what is creative?* Creativity Research Journal, 2012. **24**(1): p. 83-91.
131. Hennessey, B.A., *The consensual assessment technique: An examination of the relationship between ratings of product and process creativity.* Creativity Research Journal, 1994. **7**(2): p. 193-208.
132. Kaufman, J.C., et al., *A comparison of expert and nonexpert raters using the consensual assessment technique.* Creativity Research Journal, 2008. **20**(2): p. 171-178.
133. Daly, S.R., et al., *Comparing Ideation Techniques for Beginning Designers.* Journal of Mechanical Design, 2016. **138**(10): p. 101108.
134. Rossiter, J.R., *The new psychometrics: Comment on Appelbaum et al.(2018).* 2018.