Proceedings of the ASME 2021
International Design Engineering Technical Conferences
and Computers and Information in Engineering Conferences
IDETC/CIE2021
August 17-20, 2021, Virtual Conference, USA

DRAFT DETC2021-xxxxx

IF A PICTURE IS WORTH 1000 WORDS, IS A WORD WORTH 1000 FEATURES FOR DESIGN METRIC ESTIMATION?

Kristen Edwards
Dept. of Mechanical Engg.
Massachusetts Institute of Technology
Cambridge, Massachusetts, 02139
kme@mit.edu

Aoran Peng
Dept. of Industrial and Manufacturing Engg.
The Pennsylvania State University
State College, Pennsylvania, 16802
axp5524@psu.edu

Scarlett R. Miller
Engineering Design and Industrial Engineering
The Pennsylvania State University
State College, Pennsylvania, 16802
shm13@psu.edu

Faez Ahmed
Dept. of Mechanical Engg.
Massachusetts Institute of Technology
Cambridge, Massachusetts, 02139
faez@mit.edu

ABSTRACT

A picture is worth a thousand words, and in design metric estimation, a word may be worth a thousand features. Pictures are awarded this worth because of their ability to encode a plethora of information. When evaluating designs, we aim to capture a range of information as well, information including usefulness, uniqueness, and novelty of a design. The subjective nature of these concepts makes their evaluation difficult. Despite this, many attempts have been made and metrics developed to do so, because design evaluation is integral to innovation and the creation of novel solutions. The most common metrics used are the consensual assessment technique (CAT) and the Shah, Varga-Hernandez, and Smith (SVS) method. While CAT is accurate and often regarded as the "gold standard," it heavily relies on using expert ratings as a basis for judgement, making CAT expensive and time consuming. Comparatively, SVS is less resource-demanding, but it is often criticized as lacking sensitivity and accuracy. We aim to take advantage of the distinct strengths of both methods through machine learning. More specifically, this study seeks to investigate the possibility of using machine learning to facilitate automated creativity assessment. The SVS method results in a text-rich dataset about a design. In this paper we utilize these textual design representations and the deep semantic relationships that words and sentences encode, to predict more desirable design metrics, including CAT metrics. We demonstrate the ability of machine learning models to predict design metrics from the design itself and SVS Survey information. We demonstrate that incorporating natural language processing (NLP) improves prediction results across all of our design metrics, and that clear distinctions in the predictability of certain metrics exist.

INTRODUCTION

A picture is often said to be worth a thousand words because of the amount of information it can transmit. A picture will capture not only the object or concept of interest, but also potentially embedded interactions with the environment and, possibly, the preferences of the picture owner [1]. The same can be said with words in design concept evaluation due to the nature of languages being complex and context-dependent. For example, the same word may have different meanings and different words may have the same meanings when used in different contexts. As a result, the amount of features used to capture the dynamic state of the word increases. This is also what makes creative design evaluations difficult. The need for creative design evaluations stems from the increased attention in research on creativity and innovation in engineering, as they are crucial in providing novel solutions to new and existing problems [2–4]. It has even been said that creativity is mankind's most valuable resource, as innovation and progress rely heavily on creativity [5]. Creativity and innovation mark an individual's ability to produce new ideas, a skill that is crucial in the production of novel technology [6–9]. As a result, there has been a surge in research that examines possible methods to boost student creative and innovative behaviors [10–12].

These methods to engage students in creativity require separate methods that assess the outcomes. This is because assessments can help identify creative individuals and ideas, as well as facilitate improvements in both [13–16]. Assessments can also serve as a way of evaluating the design metrics in terms of their effectiveness at aiding the process of creative idea generation [17]. A lack in assessment strategy is not only a setback to assessing the suitability and effectiveness of these creativity-enhancing techniques in specific projects, but also may bring into question their overall effectiveness [9]. Therefore, there is a need to measure not only if a concept is creative, but also to what degree it is creative [13].

There are a plethora of metrics that aim to measure creativity today [5, 17, 18]. These metrics include, but are not limited to, expert panels [19–23], the Consensual Assessment Technique (CAT) [6, 24, 25], the Shah, Vargas-Hernandez, and Smith (SVS) method [17], and the Comparative Creativity Assessment (CCA), which is built upon the SVS method [26]. Among all of the metrics created, the most common are the CAT [6, 24, 25] and the SVS method [17]. Despite the fact that many metrics exist, measuring creativity is still difficult. One of the reasons could be due to the multi-faceted nature of creativity and what it entails [27]. In addition, the unique characteristics of the measurement methods can also result in increased challenges for researchers to establish an assessment standard [9, 28, 29]. The abundance of metrics available has resulted in great variability between the methodology of different studies, which makes comparing findings increasingly difficult [28]. For example, SVS results have been found to not match expert ratings in design variability [30], while the heavy dependence of CAT assessments on the experience, number, and subjectivity of the experts can result in the experiment being significantly restricted by their time and financial budget [18].

To address these limitations with the creativity assessment metrics, this study tries to uncover how machine learning methods can enable automated assessment of creativity. Specifically, we investigate how regression models can be used to predict the CAT ratings for unseen designs using SVS features. A total of five design metrics relating to creativity will be measured, including creativity itself. Creativity can be defined as a measure of the capacity to generate original work that is useful [6–8]. At the beginning of the rating process, ideas will be picked out from the group that represent high, medium, and low creativity [31]. The expert raters will then be trained using these anchoring concepts to evaluate other concepts by comparing them to the anchoring concepts [32]. This comparative evaluation will be captured through a 7-point Likert scale [32], and can be an accurate assessment of how relatively creative each concept is. Then, the expert will examine the concepts further in terms of its usefulness (quality and utility of the idea), uniqueness (originality of the idea), and elegance (well-crafted), which are CAT submetrics [33]. It will also be used to evaluate the drawing of the concepts as it has been found to be correlated with design outcome [34].

In this study, we will further show how Natural Language Processing (NLP) based models, which capture semantic relationships between words, can help overcome the issue of attribute dependencies in SVS features and improve the prediction results. NLP is a subset of speech and language processing that aims to train the computer to interpret the text in a more naturalistic, human way [35]. Applications of NLP include text sentiment detection and response generation [35], all of which could be helpful in completing the goal of this study. More specifically, this study can help to examine the plausibility of using SVS ratings to predict and produce CAT ratings; CAT ratings are very resource demanding, whereas utilizing SVS ratings would be faster, easier, and cheaper to gather.

RELATED WORKS

Creativity and innovation are traits that are greatly valued in the current market, as they are crucial in the formation of new ideas [2–4]. As a result, there has been an abundance of research on methods that aim to encourage and improve creativity in students [10–12]. An important step following the promotion of creativity is the assessment of creativity, which is important in the identification and evaluation of progress [13–17]. The two most commonly used metrics for creativity measurement are the CAT [6, 24, 25] and the SVS method [17]. Although the CAT method is valued as more accurate in the measurement process, it is also very resource consuming [18, 36, 37] due to the time needed to have expert raters code hundreds or thousands of ideas, plus the time needed to train novice raters if experts are unavailable. By comparison, the SVS method is faster and cheaper, as novice raters can achieve high levels of inter-rater agreement. However, it lacks sensitivity and accuracy, and does not match the rating of expert reviewers [30]. Therefore, this study was constructed to investigate the plausibility of using SVS data to predict and produce CAT ratings that accurately assess the creativity of the concepts.

Creativity Assessment Methods

Creative assessment methods are a crucial after-step of engineering design research, and are often seen as a ”means to an end” to researchers studying engineering design creativity. They are an effective tool in identifying and evaluating the effectiveness of creativity-enhancing techniques, and helping assess the novelty of ideas generated [13–17]. There have been many metrics designed that aim to measure the creativity of concepts [38] or aspects related to creativity. Research has classified these assessment methods into two categories: process-based and outcome-based [39]. More specifically, process-based methods focus on the cognitive processes that are involved during the concept generation process [39]. Comparatively, outcome-based methods examine the outcomes of the ideation process [17,39]. Of these two categories, the outcome-based methods are more commonly known and used [39]. This is primarily due to the complexity and difficulty associated with process-based approaches [17, 39]. Although research has been done to classify the different methods, there has not been one method that is denoted as the standardized method [38]. This is because each of the metrics available have their respective advantages and disadvantages. For example, although CAT has been praised as the “gold standard of creativity assessment,” [40] it is still flawed because its methods are very time and resource consuming [18, 36, 37]. Therefore, it is up to the researchers to determine, based on their needs and their resources, which method their study will employ. For this study, the focus will primarily be on CAT and SVS, two of the most commonly used metrics.

The CAT was first conceptualized and developed by Teresa Amabile to assess creativity in a subjective manner [2,24,28,41], and is often regarded as the best rating method [18, 36, 40, 42]. It measures creativity by employing a panel with appropriate expertise in the field of interest and asking them to provide their own ratings on the products or ideas generated based on a Likert scale [6, 18, 24, 25, 28, 40, 42, 43]. The process for attaining CAT ratings is as follows: (1) a group of creative concepts are gathered [18,28,40,42] and (2) raters are then asked to provide ratings from 1 (low in a factor) to 7 (high in a factor) based on the definition of each factor [28]. During the assessment, it is stressed that the experts should make their assessments independently, subjectively, and take into consideration other products under review [18,28,40,42].

The basis for this method is that an idea is only creative to the extent that experts agree, independently, that it is creative. [28,36]. Because of this, the accuracy of these assessments heavily depends on the expertise of the reviewer in that field [24,44]. This is supported by previous research that found expert ratings to have a higher agreement (higher inter-rater reliability) than non-expert ratings [18, 24, 44, 45]. Therefore, it could be said that there is no standard scoring system available for CAT, as it is entirely based on subjective comparison within a certain group [37]. This is because concepts designed in the same environment, same situation, and same predisposition can be evaluated with respect to each other [32].

Although CAT ratings are accurate because they are based on the opinion of experts in the field, they are also relatively difficult to gather [18]. Experts are often very expensive, relatively hard to find, and extremely busy [18]. This aspect of CAT makes it more difficult to implement than some of the other metrics. Human raters have also been shown to be inconsistent between each other, which can be a result of different expertise levels, as well as differences in their beliefs about creativity [46]. These are among the reasons why current researches are looking into alternatives, such as using novice raters [45], quasi-novice raters [47,48], or in the case of this paper, machine learning.

Another common metric that is used is the SVS method [17]. This method is an example of a model using the genealogical tree approach [4] and is more commonly used and accepted in engineering [17, 49]. Here, the focus is more on using effectiveness to quantify creativity of ideation [28]. When using the SVS approach, the role of the human rater is replaced by predefined components as an attempt to increase repeatability and reduce subjectivity [50]. This type of metric usually breaks down the concepts into components, and quantitatively measures the creativity of each component based on relative frequencies [17,28,51]. More specifically, the concepts will be broken down based on the function of the components [28,29]. However, this metric has been criticized as lacking in its sensitivity and accuracy [50]. For example, Linsey’s research reported that SVS results were inconsistent with the ratings produced by experts in terms of variety of concepts [30]. Other studies have reported that SVS results can have decreased accuracy as a result of an increase in sample size [52,53]. In addition, one study by Sluis-Tieschеffer *et al.* [52] found the SVS approach was unable to provide comparison between different attributes, only allowing for comparison between the same attribute.

SVS encompasses four sub-metrics: (1) novelty, (2) variety, (3) quality, and (4) quantity of ideation [17]. In this case, novelty can be defined as how different the concept is from other concepts; variety is how different the concept is from other concepts generated by the same designer; quality is a subjective measure of feasibility and degree of success at meeting desired requirement; and quantity is the number of concepts generated [17, 28, 39]. Of these four sub-metrics, quality and novelty are usually the more focused factors, as novel and appropriateness of ideas being part of the definition of creativity [6,28]. The novelty component of SVS examines how similar the idea is with other ideas from the same group [17,28,39]. Through the genealogical tree, also known as the feature tree approach, SVS proposes that novelty can be calculated based on the type of features the concept includes, as well as how each feature is satisfied [17]. Therefore, concepts that have features in categories with lower overall frequency will indicate that not many other concepts share their idea, which would then indicate higher novelty for that idea [28].

Comparatively, the quality component of SVS measures the feasibility of the concepts in terms of how successful they are at meeting the desired design requirements [17,28,39].

Machine Learning and Creativity Assessment

Data-driving research approaches and methods, like machine learning, are advantageous for analyzing large amounts of data for meaning, patterns, relationships, and even development and formation of theories [54]. In recent years, machine learning algorithms can even be used to construct models from data that are not necessarily linearly related [54]. This characteristic is what gives machine learning the potential to be used in the analysis of subjective measures. This is supported by prior research, where machine learning has been successfully used to predict subjective measures, such as mental workload [54]. It has also been used in objective quality-assessments of videos that would otherwise demand a much larger-scale and a more expensive experiment [55]. The result of this study shows that machine learning was able to significantly reduce the amount of work needed without compromising the conclusion [55]. Machine learning also shows promise in contributing to opinion-based data-mining [56]. It has already been found that opinion mining can be used to determine whether a sentence or a document is expressing positive or negative sentiment [56]. Machine learning and opinion mining has also been used with natural language processing to assess online reviews [**?**,56,57], where opinion spam detection would be used to separate out reviews while usefulness measurements [56,57] can be used to identify the usefulness and subjectivity of a review [56,58–61]. Therefore, the ability of machine learning to assist in the assessment of subjective measures, like creativity, is investigated in this study. More specifically, this study focuses on using the natural language processing.

NLP is a type of machine learning process that attempts to teach computers to interpret language in a natural, "human," way [35]. More specifically, it seeks to explore how computers can be used to understand and manipulate natural language to accomplish useful tasks [62]. Therefore, it could be said that the ultimate goal of NLP is to achieve "human-like language processing" abilities [63]. NLP has been found to be helpful in the completion of tasks like understanding the sentiment behind text and generating responses to questions [35]. For example, NLP has been used to extract information from narrative text in the medical field, and has found the results to be reasonable [64]. In another study by Li et al., NLP was used to assess Chinese subjective answers [65]. Design researchers have used NLP based methods on a plethora of tasks ranging from identifying manager interventions [66], predicting contest winners [67], sentiment analysis of conversations [68] to understanding product reviews [69]. In this paper, we show how these models, when combined with design feature information, can also enable prediction of multiple design metrics.

METHODOLOGY

Our goal is to predict expert CAT ratings of a design using SVS features and a description gathered from that design. We experiment with different methods to predict the five different CAT ratings: Usefulness, Elegance, Drawing, Uniqueness, and Creativity. Our different methods stem from using three different representations of a design, as well as three different regression models. The different design representations are derived from data available in the design itself and the unprocessed SVS features. The three design representations are:

1. One-hot-encoded SVS Features
2. One-hot-encoded SVS Features + Text Embedded Description
3. Text Embedded SVS Features and Description

Figure 1 shows how we created each of the three design representations and we will discuss this process in detail in the following sections. In addition to the three distinct design representations, we explore the use of three different regression models:

1. Linear regression
2. Gradient Boosting (GB) Regression
3. Random Forest (RF) Regression

As shown in Figure 1, we originally start with a design. Initially, we have both a text description and a completed SVS Survey about the design. In the following sections, we will discuss our methods of processing the data, our motivation for converting data into different design representations, which design representations provided the most predictive power, and how this varied for different expert CAT ratings.

SVS Data Processing

The original SVS data comes from a survey in which people are shown a design of a milk frother and asked questions regarding it, such as: "How is the device powered?" There are 91 questions total; each question and its respective response serves as an initial feature for the design. An example of two of these survey questions is shown in Figure 1, labelled by SVS survey. As the figure shows, the survey responses are categorical. Therefore, the responses have been pre-processed such that each distinct response to a question is given a number. For example, for the question "How is the device powered?" the provided response options are (1) Manually powered, (2) Electric, or (3) Other, and the corresponding number represents the response.

For example, Figure 1 shows that the responses for the two questions can be mapped to a two dimensional vector, where the first value of "3" corresponds to the third category "Other" and the second value "1" corresponds to the first category "Yes" for

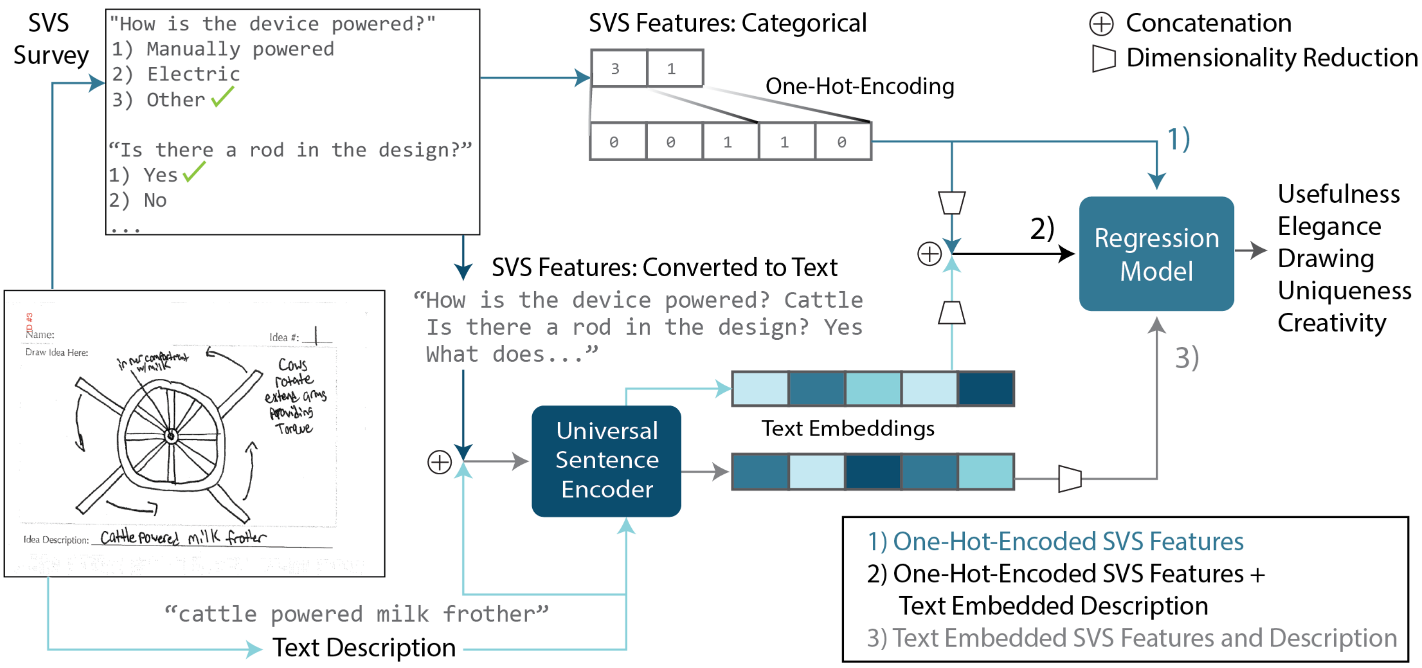
SVS Survey
"How is the device powered?"
1) Manually powered
2) Electric
3) Other ✓
"Is there a rod in the design?"
1) Yes ✓
2) No
...
SVS Features: Categorical
3
1
One-Hot-Encoding
0
0
1
1
0
⊕ Concatenation
Dimensionality Reduction
1)
2)
3)
Regression Model
Usefulness
Elegance
Drawing
Uniqueness
Creativity
SVS Features: Converted to Text
"How is the device powered? Cattle
Is there a rod in the design? Yes
What does..."
Name:
Idea #: 1
Draw Idea Here:
in our contraption w/milk
Cows rotate extend arms providing Torque
Idea Description: Cattle Powered milk frother
Universal Sentence Encoder
Text Embeddings
"cattle powered milk frother"
Text Description
1) One-Hot-Encoded SVS Features
2) One-Hot-Encoded SVS Features + Text Embedded Description
3) Text Embedded SVS Features and Description

FIGURE 1. The overall architecture of our model. From a design, we initially have a dataset of SVS features in numerical form as well as a written text description provided by the designer. We convert the numerical SVS features into text and combine that text with the original description to gain an all-text representation of the design. We encode this text representation in Tensorflow's Universal Sentence Encoder to gain a numerical text embedding for each design. We input this text embedding into a regression model to predict five expert CAT ratings: Usefulness, Elegance, Drawing, Uniqueness, and Creativity.

the second question. Although the responses were given numerical values, their relationships were not numerical, i.e. the difference between"1" and "3" is not necessarily greater than the difference between "1" and "2." Consequently, we converted these categorical features through one-hot-encoding into vectors.

One-hot-encoding is a method for converting categorical data into numerical data that can be used for a machine learning model. Categorical data often contains labels rather than numbers, for example, an image might be labeled "cat" or "dog". One-hot-encoding converts these labels into binary features. If the variable were "Type of pet" and the options were "cat" and "dog," then "cat" can be given a value [0,1] and "dog" can be given a value [1,0].

In the example shown in Figure 1, the first question had three possible answers, designated "1," "2," and "3." This single question became three binary questions, where response "1" is now [1,0,0], response "2" is now [0,1,0], and response "3" is now [0,0,1]. Each value is considered a feature, so this overall process increased the number of features from the initial 91 to 522. These 522 dimensional one-hot-encoded SVS features serve as our first design representation.

Converting SVS values to text embeddings

We were motivated to convert our SVS values into text embeddings in order to utilize the relationships between certain features. In our first design representation, all of the features are one-hot-encoded.

One-hot-encoding creates features that are all assumed to be completely independent of one another, so a design that is labeled "bolt" is as far away from a design labeled "nail," as it is from a design labeled "hammer." However, humans recognize that the first two words are semantically closer to each other than the first and third word. In this sense, one-hot-encoding all of our features forced us to lose many relationships between designs.

To address this, we converted all information about a design into text and we combined all of the text such that each design was defined by a long text string. We demonstrate this process in Figure 1. Our method used the original SVS survey responses in numerical form. For each question that had a response other than "No," we included both the question and the response as text in the representation of a design. For example, for the second question in Figure 1, "Is there a rod in the design?" the response is "Yes," therefore, the question and the response are both included in the string of text that represents the design.

For the first question in Figure 1, "How is the device powered?," the response is "Other." Whenever a respondent selected

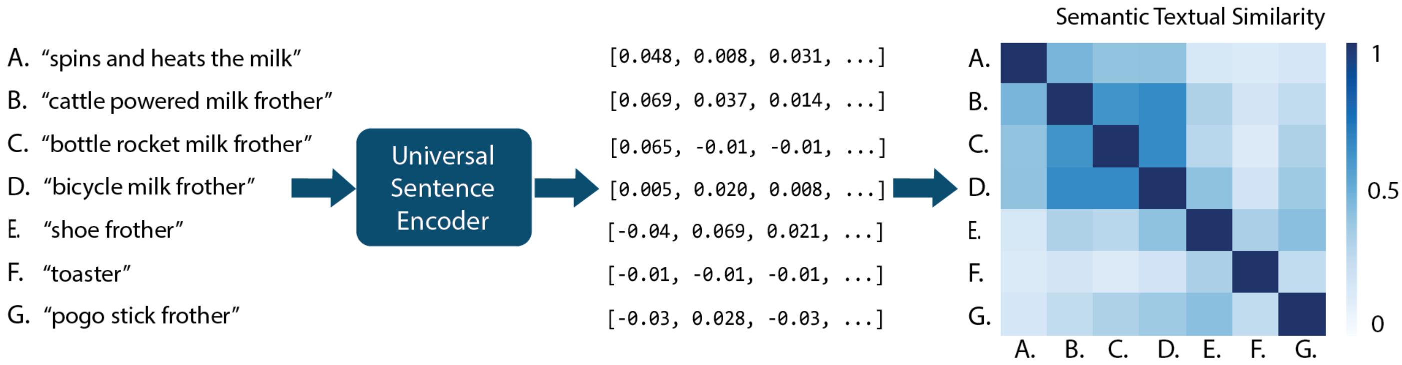

FIGURE 2. The process of generating text embedding using Tensorflow's Universal Sentence Encoder. Shown on the left, the text descriptions for each design are inputted into the autoencoder. The 1 x 512 numerical embeddings for each description are found, and a heat map of the semantic textual similarity is shown on the right, where a score of 1 indicates maximum similarity.

"Other," the respondents provided a description themselves. In this case, the description was "Cattle," so both the question "How is the device powered?" and the response "Cattle" are included in the string of text that represents the design.

Lastly, we add the original text description to the text string representation. With this new design representation, all of the features are in text format. The next step is to convert this textual design representation into numerical continuous embeddings that can be used in regression models.

Universal Sentence Encoder

Figure 1 shows that we run both the text description and the SVS features as text through a sentence encoder in order to get text embeddings. A text embedding is a vector representation of text in which text with similar meanings are represented with similar vectors.

We use Tensorflow's Universal Sentence Encoder, which maps text into 512-dimensional space [70]. Figure 2 demonstrates how we use the Universal Sentence Encoder. We input a string of text such as "spins and heats the milk," then the Universal Sentence Encoder maps the text into 512-dimensional text and outputs a text embedding: a 512-dimensional vector of numbers that represents the inputted text.

The Semantic Textual Similarity plot shown in Figure 2 illustrates the effectiveness of the Universal Sentence Encoder. In the plot, the original sentences are represented with letters A-G. After calculating the text embeddings, we use cosine vector similarity to find which sentences are more similar. The similarities between the seven sentences are shown by the 7x7 matrix at the right. A darker square indicates a higher similarity between two sentences. For example, sentences B-D all include the word "milk frother," and these sentences are shown to have some of the most similarity (darkest squares). In contrast, sentences E-G are quite different from each other and consequently have low similarity scores and light squares. We also notice deeper relationships being captured by the Universal Sentence Encoder. For instance, the most similar item in this list to "pogo stick frother" is a "shoe frother," which relates to the deeper connection between a "shoe" and a "pogo stick."

Dimensionality Reduction

In the above sections, we discussed using both one-hot-encoding and text embeddings to represent our designs. Both of these methods output vectors of over 500 dimensions, or features. Because we have 934 designs (each representing a datapoint for a machine learning algorithm), which further reduces by 20% after the train-test split, we face the "curse of dimensionality." We have too many features for our datapoints, and we need to reduce the number of dimensions in our feature space.

We utilize principal component analysis (PCA) to reduce the dimensions of our feature space. PCA is a tool from linear algebra that projects our original features into a lower-dimensional space. PCA accomplishes this by taking linear combinations of the original features, thereby creating new features. These new features are principal components, and they are independent of one another and also retain most of the information from the original features. PCA aims to put the most information possible in the first component, and the second most in the second component, and so on. Finally, PCA drops the principal components that have the least information, thereby reducing the number of features we have.

Design Representations

The three different design representations come from applying various combinations of the data processing methods mentioned above. Ultimately, the first design representation is a one-hot-encoding of all of the SVS features - resulting in 522 features total. We illustrate the process of generating this representation in Figure 1, where it is denoted as "1."

The second design representation combines one-hot-encoding, text embedding, and PCA, as illustrated in Figure 1. This design representation results in 30 features total, 15 from the one-hot-encoded SVS features after PCA, and 15 from the embeddings of the text descriptions after PCA.

The third design representation converts all of the information available about a design into text - the SVS data as well as the text descriptions. From that complete text representation, the third design representation finds a text embedding. This text embedding is reduced via PCA from 512 dimensions to 30, such that the third design representation includes only 30 features.

Regression Models and Evaluation Metrics

We experimented with generating CAT rating predictions with three different regression models: linear regression, gradient boosting (GB) regression, and random forest (RF) regression.

Linear regression attempts to model the relationship between independent variables and a dependent variable through a linear equation. We train the linear regression by minimizing the sum of squared differences between our predicted and actual values. This method is known as the least-squares method.

The gradient boosting and random forest regression models are both ensembles of decision trees. An ensemble means that these models are aggregations of other models. The motivation behind an ensemble method is to combine predictions from multiple base models into a prediction that is better than that of any single model [71]. GB trains multiple decision trees sequentially, which is an ensemble method called boosting. The first decision tree is able to predict the majority of the data, and the following trees work to capture areas of the data that have been missed.

RF utilizes bagging, or training individual models in parallel on a randomized subset of the data, as the ensemble method [71]. RF takes the average of the predictions. This combination of decision trees yields better predictions because it has lower variance compared to a single decision tree. We train both GB and RF using the least-squares method.

We perform supervised learning with an 80-20 train-test split for each of our models. We have 934 designs total, so we train using 747 designs, and test using the remaining 187.

We use the R^2 value as the evaluation metric of our regression models. For each model we predict CAT ratings of the designs in the test set. We compare these predictions to the actual CAT ratings and use this comparison as a means for evaluating how effective our model is. The R^2 value, or coefficient of determination, quantifies the degree to which our predicted values are linearly correlated with our actual values. A perfect R^2 score is 1, indicating perfect linear correlation, and an R^2 score of 0.3 is generally accepted as indicating a weak positive correlation.

RESULTS

In this section, we discuss how effectively we predicted CAT expert ratings, with three varying parameters: (1) which CAT rating we are predicting, (2) which design representation we are inputting, and (3) which regression model we are using. We found that all three of these parameters impact the effectiveness of our predictions, as evaluated by the R^2 metric. The following sections are divided based on design representation, and, within each section, we explore results for predicting each CAT rating with each regression model.

Predicting design metrics from One-Hot-Encoded SVS Features

The left three columns of Table 1 shows the R^2 score obtained using the first design representation: one-hot-encoded SVS features. Overall, linear regression performs the worst, with consistently large and negative R^2 scores. Gradient boosting and random forest regressions perform comparably, with gradient boosting slightly outperforming random forest for every CAT metric. The trends along the design metrics are also distinct. Due to the poor nature of the linear regression results in this particular experiment, we will only discuss trends seen within the gradient boosting and random forest regressions. For the GB and RF models, the Usefulness design metric shown in the final row of Table 1 has the highest R^2 value, with Elegance following behind. Creativity and Uniqueness show some of the worst R^2 scores, Creativity being the worst with negative R^2 values across both the GB and RF models.

Converting SVS values to text embeddings

Figure 5 demonstrates the motivation and effectiveness of using NLP in this context. The figure displays three designs for milk frothers, from left to right: a Bicycle design, a Cattle design, and a Rodeo design. The Bicycle design involves attaching containers of milk to the spokes of a bicycle wheel and frothing the milk through the motion of riding a bike. The Cattle design involves putting milk inside a horizontal wheel, cattle will push the spokes of the wheel as they walk in circles, which froths the milk. The Rodeo design involves attaching a container of milk to the back of a mechanical bull, and the motion of riding that bull will froth milk. For context, many other designs in the dataset are a variation of whipping milk with a whisk, so these three designs are qualitatively unique and similar to each other.

Our goal is to create an objective model (as opposed to subjective expert ratings) that captures relationships among designs in order to more effectively predict their CAT ratings. The results shown in Figure 5 illustrate that using NLP in design representation captures these relationships much more effectively than using discontinuous one-hot-encoding representations.

The bar chart shown in Figure 5 displays the cosine similarity between designs for two design representations: One-Hot-

	One-hot encoded SVS features			One-hot encoded SVS features + Description text embeddings			Text-based SVS features + Description text embeddings		
	Linear Regression	**Gradient Boosting**	**Random Forest**	**Linear Regression**	**Gradient Boosting**	**Random Forest**	**Linear Regression**	**Gradient Boosting**	**Random Forest**
Creativity	-9.54E+25	-0.033	-0.263	-0.023	-0.114	-0.098	-0.014	-0.085	**0.034**
Uniqueness	-7.24E+25	**0.070**	-0.020	-0.109	-0.099	-0.043	0.016	-0.159	0.057
Drawing	-5.57E+24	0.117	-0.031	**0.163**	0.114	0.113	0.150	0.064	0.099
Elegance	-2.53E+25	0.162	0.100	0.140	0.058	0.139	0.179	0.079	**0.211**
Usefulness	-4.05E+26	0.171	0.162	0.149	0.131	0.214	0.209	0.180	**0.263**

TABLE 1. R^2 Score by Regression Model using SVS One-Hot-Encodings as the design representation

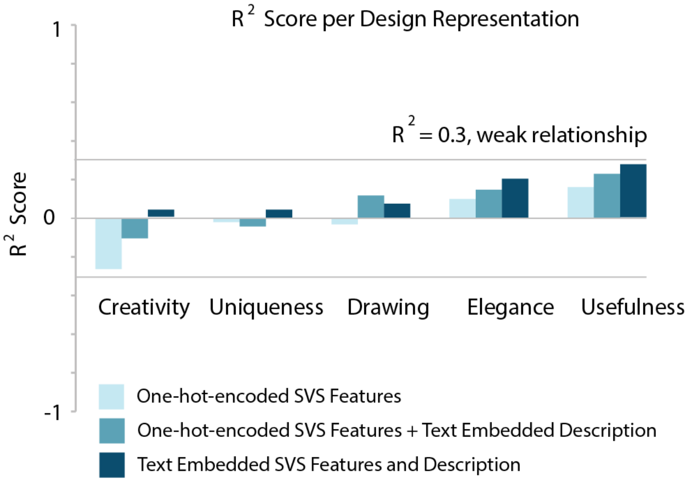

FIGURE 3. This graph shows the R^2 score for each of the five CAT metrics with three different design representations (1) one-hot-encoded SVS features, (2) one-hot-encoded SVS Features as well as text embedded descriptions, (3) text embedded SVS features and descriptions. The line marked at 0.3 indicates an R^2 score threshold of a weak positive relationship. These results are generated with the random forest regression model. We observe that, overall, adding text embeddings improves the regression results, as seen by the increase in R^2 scores. The figure highlights that despite the improvements, the regression results for all the design metrics are still below the threshold of weak relationship.

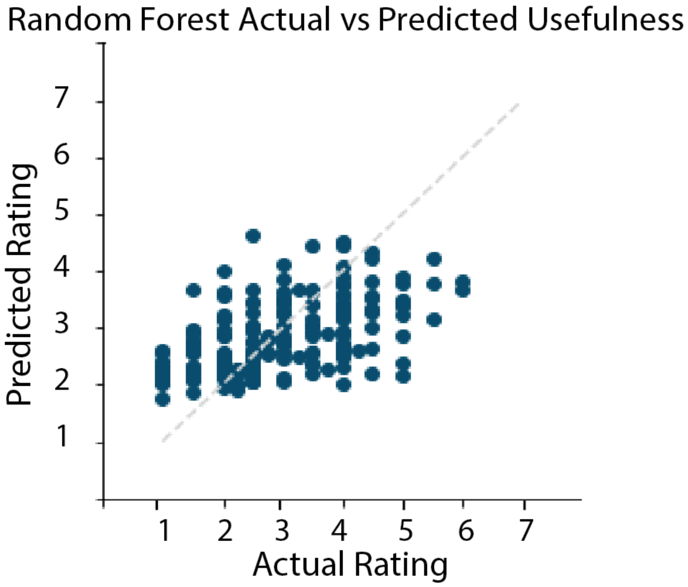

FIGURE 4. A scatterplot of the predicted Usefulness rating vs. the actual Usefulness rating found using a Random Forest Regression Model. A perfect prediction would follow a line with a slope of one and intercept at the origin, represented in the plot by a black line. This prediction has an R^2 score of 0.270.

Encoded SVS Features, and Text Embedded SVS Features and Description, which are generated using NLP. Cosine similarity is a measure of similarity between two vectors, found by calculating their inner product. Cosine similarities range between 0 to 1 and for two vectors A and B, it is defined as: $Sim(A,B) = \frac{A \cdot B}{\|A\|\|B\|}$

In the example of the Bicycle and Rodeo designs, the cosine similarity between their one-hot-encoded representations is 0.333, whereas the cosine similarity between the text embedded representations is 0.810. This may suggest that the text embedded representations maintained relationships between the designs that are lost when using one-hot-encoded representations. These results comply with our understanding of the two types of representation. One-hot-encoded vectors ensure that each feature has a cosine similarity of 0 with any other feature, since each feature that exists has a value of 0 in all dimensions except

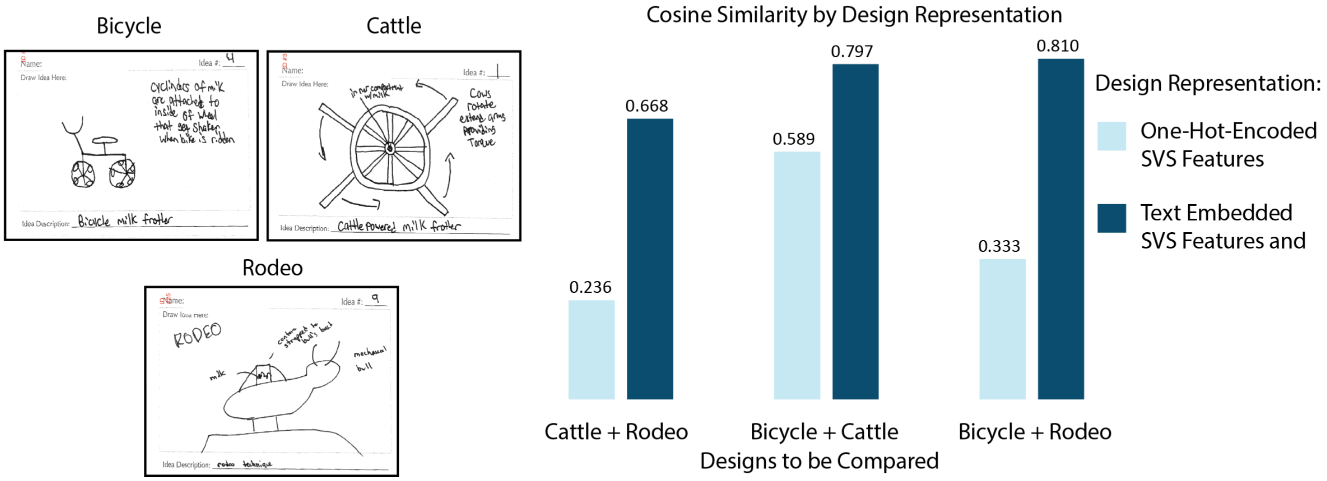

FIGURE 5. The top portion shows three student-generated designs for milk frothers, a Bicycle design, Cattle design, and Rodeo design. Under each design is a handwritten description provided by the designer. We measure the similarity of these designs by finding the cosine similarity of their representations. We use two different representation methods: one-hot-encoding of their SVS features and text embeddings of their SVS features and descriptions. Note that the similarities using one-hot-encoded features between Bicycle & Rodeo is lesser than the similarity between Bicycle & Cattle. However, text embeddings highlight their semantic similarity where riding is involved in both the devices to froth the milk.

		Actual Class		
		Low	Medium	High
Predicted Class	Low	39	20	7
	Medium	13	24	15
	High	11	18	40

TABLE 2. Confusion matrix for classifying the Usefulness of a design using a random forest regression and percentile classification. The precision values for the low, medium, and high classes are 0.619, 0.387, and 0.645 respectively. The recall values for the low, medium, and high classes were 0.591, 0.462, and 0.580 respectively.

one, where it has a value of 1. Each one-hot-encoded vector of a feature is orthogonal to all others. In contrast, text embeddings found through NLP, such as those we used through Tensorflow's Universal Sentence Encoder, map all of the features in a continuous space, which preserves deeper relationships between them.

Predicting design metrics from One-Hot-Encoded SVS features + Text Embedded Description

Our second design representation utilizes NLP and represents designs using one-hot-encoded SVS features as well as a text embedding of the design's description. As described in the Design Representations section of Methodology, we perform PCA dimensionality reduction on both of these vectors, resulting in 30 total features for the second design representation.

The middle three columns of Table 1 shows the R^2 scores for each of the CAT ratings with predictions from each of the three regression models. As compared to Table 1, the linear regression model performs significantly better with this design representation. All three regression models perform comparably. The performance of the CAT ratings matches the general trend seen in the first design representation. Usefulness and Elegance are the best predicted. With the new design representation, Drawing is more accurately predicted in both linear regression and random forest, while gradient boosting's prediction is comparable to its prediction from the first design representation.

The overall trend across different design representations is illustrated in Figure 3 for the random forest regression. The figure indicates an overall improvement in R^2 scores from the first design representation: One-Hot-Encoded SVS Features, to the second design representation: One-Hot-Encoded SVS Features + Text Embedded Description.

Predicting design metrics from Text Embedded SVS Features and Description

Our third design representation is a text embedding of both SVS features and the text description. All of the information from the SVS Features is converted to text, combined with the original text description, mapped to a continuous space, and then encoded as text embeddings that preserve the semantic relationships from the continuous mapping.

Regression results: The results from using this design representation are shown in the right three columns of Table 1, which shows the R^2 score found for each combination of CAT rating and regression model. The results are consistent with the overall trend seen in both design representation one and two. Usefulness and Elegance are the CAT ratings that are best predicted, as demonstrated by the highest R^2 scores across all three regression models. For both of these metrics, all three regression models perform the better with this NLP based design representation than with the other two design representations. Additionally, for both the linear regression and random forest models, the performance of Usefulness and Elegance have consistently improved from design representation one to design representation three. This overall trend can be observed in Figure 3 for the random forest model.

In agreement with the trend we have observed in the other two design representations, the Creativity and Uniqueness metrics have the lowest R^2 scores. We will discuss our hypotheses explaining why certain CAT metrics are consistently more or less accurately predicted in the Discussion section to follow.

We show a visual representation of the predicted CAT ratings vs. the actual CAT ratings in Figure 4. This scatter plot is generated using predictions from the random forest regression model using the third design representation: Text Embedded SVS Features and Description, and the Usefulness CAT rating. A set of perfect predictions would follow the black line with a one-to-one slope between the Actual and Predicted Ratings.

The plot reveals that most of the Predicted Ratings range between 2 and 4. This trend appears for other CAT ratings as well. Remedying the models to more successfully predict CAT ratings on the high and low extremes, perhaps through increasing the weight of designs with extreme Actual Ratings, could be an area of future exploration.

Classification results: In response to the low R^2 values we found with a regression model, we tested how effective our model is at predicting the relative class of a design metric. Our classes are determined by the percentile of a design's rating with regard to all other designs in a test set. Designs with ratings between the 1-33 percentiles are in the low class, designs with ratings between the 34-67 percentiles are in the medium class, and designs with ratings between the 67-100 percentiles are in the high class.

The results are demonstrated in the confusion matrix in Table 2. The confusion matrix illustrates agreement between predicted and actual classes. A perfect classification has numbers only along the diagonal and zeroes elsewhere. Not all errors are equal in confusion matrices. Predicting a high value when the true value is low (or vice versa) is an extreme error, which we aim to minimize the most.

In Table 2, we see that we predict a low class when the real class is high 7 times, and do the opposite 11 times. This compares to correctly predicting the low class 39 times and the high class 40 times.

We also observe the precision metric, which indicates how many positive predictions are true. It is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

For the low, medium, and high classes the precision values are 0.619, 0.387, and 0.645 respectively.

Recall, also known as the true positive rate (TPR), measures how many of the positive cases our model is able to correctly predict. Recall is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

For the low, medium, and high classes the recall values were 0.591, 0.462, and 0.580 respectively. The high recall values show that our regression model is able to effectively classify designs into low, medium and high categories, which can enable human raters to use these models for initial filtering of ideas and then focusing on individual categories to identify the the top ideas.

DISCUSSION AND LIMITATIONS

The goal of this study was to identify how to take advantage of the distinct strengths of both the SVS and CAT method through machine learning. More specifically, this study sought to investigate the possibility of using machine learning to facilitate automated creativity assessment. Our results revealed two major and consistent trends:

1. Incorporating natural language processing in the representation of a design consistently improves a model's ability to predict expert design metrics.
2. Different design metrics vary distinctly in their predictability, with Usefulness and Elegance performing the best, and Creativity and Uniqueness performing the worst.

Trend 1 is displayed visually in Figure 3. Each design representation, from one to three, incorporates more text added. On the figure this is visually displayed with the lightest bar incorporating the least NLP and the darkest bar incorporating the most. Across all design metrics a general trend emerges. In design representation three, Text Embedded SVS Features and Description tends to outperform both other design representations, while design representation two tends to outperform design representation one.

We propose that this trend is due to Universal Sentence Encoder's ability to capture relationships between features by mapping them in a continuous space. Universal Sentence Encoder's ability to do this is in contrast to one-hot-encoding vectorizations, which assume all features are independent. We note that the overall R^2 values are low. Our predictions have much room for improvement, as visualized in Figure 4. However, the emerging trends still provide insight that we find valuable in how we can predict classically subjective ratings of designs objectively.

The other emerging trend that we found compelling is the consistent difference in the ability to predict certain CAT ratings. We found that Usefulness and Elegance metrics were consistently the most predictable. This trend could result from the objective nature of Usefulness and Elegance. More raters may agree on what characteristics of a design qualifies as useful. Furthermore, elegance is tied to the simplicity of a design, which can be more objectively agreed upon than, say, creativity. To that note, Creativity and Uniqueness were the least predictable across all regression models. We attribute this to the subjectivity of these ratings. In fact, Creativity is often not agreed upon by experts, and expert ratings are the basis of our regression models.

This finding opens another discussion centered around the availability of expert vs. novice design ratings. Novice design ratings are less expensive and more easily acquired than expert design ratings. However, expert design ratings naturally hold more weight - novices must be trained to be 'expert-proxies' using some sample set from expert ratings. Even when you can train novices to be 'expert-like,' they still lack the mental models and experiences of experts which ultimately impact rating performance. However, some areas do exist where experts and novices tend to agree such as in the novelty of design ratings [28]; thus, one could argue that there are instances where novice ratings can supplement or replace expert ones.

Future work will focus on combining language features with sketch understanding and computer vision methods for improved prediction of design metrics. We believe that a sketch contains rich information, which may get lost when a rater tries to identify SVS features from it or the designer tries to write a description. Another direction of research will be training models to identify design features from the sketch and text description. Such a model can augment our understanding of sketches and help in automated categorization and assessment of design sketches.

CONCLUSION

Creativity and innovation are important steps in the development of novel solutions to existing and new problems and for making important technological progress [2–4]. As a result, both creativity and innovation have often been viewed to be man-kind's most valuable resources [5]. As a result, there have been many attempts to help boost creativity of students to better prepare them for the markets [10–12]. One important step that comes after the implementation of improvement methods is assessment, since it can help identify whether progress has been made [13–17].

The need for creative design evaluations stems from the increased attention in research on creativity and innovation in engineering, as they are crucial in providing novel solutions to new and existing problems [2–4]. Creativity and innovation mark an individual's ability to produce new ideas, something that is crucial in the production of novel technology [6–9]. As a result, there has been a surge in research that examines possible methods to boost student creative and innovative behaviors [10–12]. Like improvement methods, there have also been a plethora of assessment metrics, the most common of which is the CAT [6, 24, 25] and the SVS method [17]. However, both techniques have their advantages and disadvantages. While CAT has been widely accepted as the "gold standard" [18, 36, 40, 42], it relies heavily on the subjective judgement of human experts in that domain [6, 18, 24, 25, 28]. This resulted in CAT being very resource consuming, both in terms of time and cost [18]. On the other hand, the SVS technique minimizes the need for experts by basing their ratings on the component-level functions of the concepts [4, 28, 29]. This makes SVS ratings faster and less expensive to produce. However, it has been criticized for lacking sensitivity and accuracy [4, 52]. Therefore, to help address this problem, this study was created to investigate the possibility of using machine learning to facilitate automated creativity assessment. More specifically, this study seeks the possibility of taking advantage of both methods by incorporating machine learning to use SVS ratings, which are easier to collect, to predict CAT ratings. This is done by using regression models in the prediction process, and also by exploring the possibility of using NLP based models to improve the results.

Our results, although preliminary, show that incorporating NLP in the prediction process can improve the model's prediction of CAT ratings. This study also found that the predictability of different aspects of the CAT ratings vary, with Usefulness and Elegance having the best predictability. These findings can serve as empirical evidence supporting the investigation of novice vs expert usage in creativity assessment. In addition, the results also can serve as empirical evidence on the plausibility of using machine learning to facilitate creativity assessment. The preliminary success in using NLP to predict CAT ratings can help to show the wide application of NLP, and also support its usage in modeling subjective ratings like creativity. In conclusion, we hope that the study can shed some light on the ongoing debate of using novices vs experts in creativity assessments. We also hope that this study can help to support future investigation by providing a direction at a new possibility of concept evaluation.

ACKNOWLEDGMENT

We thank the Ida M. Green Fellowship for supporting Kristen Edwards's research.

REFERENCES

[1] Snyder, J., 2014. "Visual representation of information as communicative practice". *Journal of the Association for Information Science and Technology,* ***65***(11), pp. 2233–2247.

[2] Amabile, T. M., 1988. "A model of creativity and innovation in organizations". *Research in organizational behavior,* ***10***(1), pp. 123–167.

[3] Maritz, A., and Donovan, J., 2015. "Entrepreneurship and innovation: Setting an agenda for greater discipline contextualisation". *Education & training (London),* ***57***(1), pp. 74–87.

[4] Ahmed, F., Ramachandran, S. K., Fuge, M., Hunter, S., and Miller, S., 2019. "Interpreting idea maps: Pairwise comparisons reveal what makes ideas novel". *Journal of Mechanical Design,* ***141***(2).

[5] Batey, M., and Furnham, A., 2006. "Creativity, intelligence, and personality: A critical review of the scattered literature". *Genetic, social, and general psychology monographs,* ***132***(4), pp. 355–429.

[6] Ambile, T. M., 1996. *Creativity in Context*. Westview Press, Boulder, Colorado.

[7] Sternberg, R. J., 1999. *Handbook of creativity*. Cambridge University Press.

[8] Mumford, M. D., and Gustafson, S. B., 1988. "Creativity syndrome: Integration, application, and innovation.". *Psychological bulletin,* ***103***(1), p. 27.

[9] Liikkanen, L. A., Hämäläinen, M. M., Häggman, A., Björklund, T., and Koskinen, M. P., 2011. "Quantitative evaluation of the effectiveness of idea generation in the wild". In International Conference on Human Centered Design, Springer, pp. 120–129.

[10] Louridas, P., 1999. "Design as bricolage: anthropology meets design thinking". *Design Studies,* ***20***(6), pp. 517–535.

[11] Toh, C. A., and Miller, S. R., 2016. "Creativity in design teams: the influence of personality traits and risk attitudes on creative concept selection". *Research in Engineering Design,* ***27***(1), pp. 73–89.

[12] Sarkar, P., and Chakrabarti, A., 2014. "Ideas generated in conceptual design and their effects on creativity". *Research in Engineering Design,* ***25***(3), pp. 185–201.

[13] Sarkar, P., and Chakrabarti, A., 2011. "Assessing design creativity". *Design studies,* ***32***(4), pp. 348–383.

[14] Sundström, P., and Zika-Viktorsson, A., 2003. "Innovation through explorative thinking in product development projects". In DS 31: Proceedings of ICED 03, the 14th International Conference on Engineering Design, Stockholm.

[15] Christensen, B. T., and Ball, L. J., 2016. "Dimensions of creative evaluation: Distinct design and reasoning strategies for aesthetic, functional and originality judgments". *Design Studies,* ***45***, pp. 116–136.

[16] Eshun, E. F., and de Graft-Johnson, K., 2012. "Learner perceptions of assessment of creative products in communication design". *Art, Design & Communication in Higher Education,* ***10***(1), pp. 89–102.

[17] Shah, J. J., Smith, S. M., and Vargas-Hernandez, N., 2003. "Metrics for measuring ideation effectiveness". *Design studies,* ***24***(2), pp. 111–134.

[18] Cseh, G. M., and Jeffries, K. K., 2019. "A scattered cat: A critical evaluation of the consensual assessment technique for creativity research.". *Psychology of Aesthetics, Creativity, and the Arts,* ***13***(2), p. 159.

[19] Alipour, L., Faizi, M., Moradi, A. M., and Akrami, G., 2017. "The impact of designers' goals on design-by-analogy". *Design Studies,* ***51***, pp. 1–24.

[20] Cheng, P., Mugge, R., and Schoormans, J. P., 2014. "A new strategy to reduce design fixation: Presenting partial photographs to designers". *Design Studies,* ***35***(4), pp. 374–391.

[21] Chan, J., Dow, S. P., and Schunn, C. D., 2018. "Do the best design ideas (really) come from conceptually distant sources of inspiration?". In *Engineering a Better Future*. Springer, Cham, pp. 111–139.

[22] Baer, J., 2015. "The importance of domain-specific expertise in creativity". *Roeper Review,* ***37***(3), pp. 165–178.

[23] Galati, F., 2015. "Complexity of judgment: What makes possible the convergence of expert and nonexpert ratings in assessing creativity". *Creativity Research Journal,* ***27***(1), pp. 24–30.

[24] Amabile, T. M., 1982. "Social psychology of creativity: A consensual assessment technique.". *Journal of personality and social psychology,* ***43***(5), p. 997.

[25] Baer, J., and Kaufman, J. C., 2019. "Assessing creativity with the consensual assessment technique". In *The Palgrave Handbook of Social Creativity Research*. Springer, pp. 27–37.

[26] Linsey, J. S., Green, M. G., Murphy, J. T., Wood, K. L., and Markman, A. B., 2005. ""collaborating to success": An experimental study of group idea generation techniques". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 4742, pp. 277–290.

[27] Ramachandran, S. K., 2019. "Investigating the accuracy of creativity metrics used in engineering design". PhD thesis, Pennsylvania State University.

[28] Miller, S. R., Hunter, S. T., Starkey, E., Ramachandran, S., Ahmed, F., and Fuge, M., 2021. "How should we measure

creativity in engineering design? a comparison between social science and engineering approaches". *Journal of Mechanical Design,* ***143***(3).

[29] Atilola, O., Tomko, M., and Linsey, J. S., 2016. "The effects of representation on idea generation and design fixation: A study comparing sketches and function trees". *Design Studies,* ***42***, pp. 110–136.

[30] Linsey, J. S., 2007. "Design-by-analogy and representation in innovative engineering concept generation". PhD thesis.

[31] Redmond, M. R., Mumford, M. D., and Teach, R., 1993. "Putting creativity to work: Effects of leader behavior on subordinate creativity". *Organizational behavior and human decision processes,* ***55***(1), pp. 120–151.

[32] Gosnell, C. A., and Miller, S. R., 2016. "But is it creative? delineating the impact of expertise and concept ratings on creative concept selection". *Journal of Mechanical Design,* ***138***(2).

[33] Besemer, S. P., 1998. "Creative product analysis matrix: testing the model structure and a comparison among products–three novel chairs". *Creativity Research Journal,* ***11***(4), pp. 333–346.

[34] Yang, M. C., 2009. "Observations on concept generation and sketching in engineering design". *Research in Engineering Design,* ***20***(1), pp. 1–11.

[35] Beysolow II, T., 2018. *What Is Natural Language Processing?* Apress, Berkeley, CA, pp. 1–12.

[36] John, B., and Sharon, S. M., 2009. *Assessing Creativity Using the Consensual Assessment Technique*. IGI Global, Hershey, PA, USA, pp. 65–77.

[37] Kaufman, J. C., Baer, J., and Cole, J. C., 2011. "Expertise, domains, and the consensual assessment technique". *The Journal of creative behavior,* ***43***(4), pp. 223–233.

[38] Johnson, T. A., Cheeley, A., Caldwell, B. W., and Green, M. G., 2016. "Comparison and extension of novelty metrics for problem-solving tasks". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 50190, American Society of Mechanical Engineers, p. V007T06A012.

[39] Nelson, B. A., Wilson, J. O., Rosen, D., and Yen, J., 2009. "Refined metrics for measuring ideation effectiveness". *Design Studies,* ***30***(6), pp. 737–743.

[40] Barth, P., and Stadtmann, G., 2020. "Creativity assessment over time: Examining the reliability of cat ratings". *The Journal of Creative Behavior*.

[41] Amabile, T. M., 2018. *Creativity in context: Update to the social psychology of creativity*. Routledge.

[42] Kaufman, J. C., Plucker, J. A., and Baer, J., 2008. *Essentials of creativity assessment*, Vol. 53. John Wiley & Sons.

[43] Baer, J., Kaufman, J. C., and Gentile, C. A., 2004. "Extension of the consensual assessment technique to nonparallel creative products". *Creativity research journal,* ***16***(1), pp. 113–117.

[44] Amabile, T. M., 1983. "Brilliant but cruel: Perceptions of negative evaluators". *Journal of Experimental Social Psychology,* ***19***(2), pp. 146–156.

[45] Kaufman, J. C., Baer, J., Cole, J. C., and Sexton, J. D., 2008. "A comparison of expert and nonexpert raters using the consensual assessment technique". *Creativity Research Journal,* ***20***(2), pp. 171–178.

[46] Long, H., and Pang, W., 2015. "Rater effects in creativity assessment: A mixed methods investigation". *Thinking Skills and Creativity,* ***15***, pp. 13–25.

[47] Kaufman, J. C., Baer, J., Cropley, D. H., Reiter-Palmon, R., and Sinnett, S., 2013. "Furious activity vs. understanding: How much expertise is needed to evaluate creative work?". *Psychology of Aesthetics, Creativity, and the Arts,* ***7***(4), p. 332.

[48] Kaufman, J. C., Gentile, C. A., and Baer, J., 2005. "Do gifted student writers and creative writing experts rate creativity the same way?". *Gifted Child Quarterly,* ***49***(3), pp. 260–265.

[49] Oman, S. K., Tumer, I. Y., Wood, K., and Seepersad, C., 2013. "A comparison of creativity and innovation metrics and sample validation through in-class design projects". *Research in Engineering Design,* ***24***(1), pp. 65–92.

[50] Ahmed, F., Ramachandran, S. K., Fuge, M., Hunter, S., and Miller, S., 2019. "Measuring and optimizing design variety using herfindahl index". In ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection.

[51] Shah, J. J., Kulkarni, S. V., and Vargas-Hernandez, N., 2000. "Evaluation of idea generation methods for conceptual design: effectiveness metrics and design of experiments". *J. Mech. Des.,* ***122***(4), pp. 377–384.

[52] Sluis-Thiescheffer, W., Bekker, T., Eggen, B., Vermeeren, A., and De Ridder, H., 2016. "Measuring and comparing novelty for design solutions generated by young children through different design methods". *Design Studies,* ***43***, pp. 48–73.

[53] Peeters, J., Verhaegen, P.-A., Vandevenne, D., and Duflou, J., 2010. "Refined metrics for measuring novelty in ideation". *IDMME Virtual Concept Research in Interaction Design, Oct*, pp. 20–22.

[54] Moustafa, K., Luz, S., and Longo, L., 2017. "Assessment of mental workload: a comparison of machine learning methods and subjective assessment techniques". In International symposium on human mental workload: Models and applications, Springer, pp. 30–50.

[55] Aldahdooh, A., Masala, E., Van Wallendael, G., Lambert, P., and Barkowsky, M., 2019. "Improving relevant subjective testing for validation: Comparing machine learning algorithms for finding similarities in vqa datasets using objective measures". *Signal Processing: Image Communica-*

tion, ***74***, pp. 32–41.

[56] Sun, S., Luo, C., and Chen, J., 2017. "A review of natural language processing techniques for opinion mining systems". *Information fusion,* ***36***, pp. 10–25.

[57] Jindal, N., and Liu, B., 2008. "Opinion spam and analysis". In Proceedings of the 2008 international conference on web search and data mining, pp. 219–230.

[58] Ghose, A., and Ipeirotis, P. G., 2007. "Designing novel review ranking systems: predicting the usefulness and impact of reviews". In Proceedings of the ninth international conference on Electronic commerce, pp. 303–310.

[59] Liu, Y., Huang, X., An, A., and Yu, X., 2008. "Modeling and predicting the helpfulness of online reviews". In 2008 Eighth IEEE international conference on data mining, IEEE, pp. 443–452.

[60] Lu, Y., Tsaparas, P., Ntoulas, A., and Polanyi, L., 2010. "Exploiting social context for review quality prediction". In Proceedings of the 19th international conference on World wide web, pp. 691–700.

[61] Ghose, A., and Ipeirotis, P. G., 2010. "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics". *IEEE transactions on knowledge and data engineering,* ***23***(10), pp. 1498–1512.

[62] Chowdhury, G. G., 2003. "Natural language processing". *Annual review of information science and technology,* ***37***(1), pp. 51–89.

[63] Liddy, E. D., 2001. "Natural language processing".

[64] Meystre, S., and Haug, P. J., 2006. "Natural language processing to extract medical problems from electronic clinical documents: performance evaluation". *Journal of biomedical informatics,* ***39***(6), pp. 589–599.

[65] Li, R., Zhu, Y., and Wu, Z., 2013. "A new algorithm to the automated assessment of the chinese subjective answer". In 2013 International Conference on Information Technology and Applications, IEEE, pp. 228–231.

[66] Gyory, J. T., Kotovsky, K., and Cagan, J., 2020. "A topic modeling approach to study the impact of manager interventions on design team cognition". In ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection.

[67] Ahmed, F., and Fuge, M., 2017. "Capturing winning ideas in online design communities". In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 1675–1687.

[68] Dehbozorgi, N., Maher, M. L., and Dorodchi, M., 2020. "Sentiment analysis on conversations in collaborative active learning as an early predictor of performance". In 2020 IEEE Frontiers in Education Conference (FIE), IEEE, pp. 1–9.

[69] Joung, J., and Kim, H. M., 2020. "Importance-performance analysis of product attributes using explainable deep neural network from online reviews". In ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection.

[70] Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R., 2018. Universal sentence encoder.

[71] Scikit-learn. *Ensemble Methods*. `https://scikit-learn.org/stable/modules/ensemble.html`.