**Proceedings of the ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference**
**IDETC/CIE 2019**
**August 18 - 21, 2019, Anaheim, CA, USA**

# DETC2019-97778

# MEASURING AND OPTIMIZING DESIGN VARIETY USING HERFINDAHL INDEX

**Faez Ahmed***
Dept. of Mechanical Engineering
University of Maryland
College Park, Maryland 20742
Email: faez00@umd.edu

**Sharath Kumar Ramachandran**
School of Engineering Design, Technology and Professional Programs
The Pennsylvania State University
University Park, PA
Email: sharath@psu.edu

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

**Sam Hunter**
Industrial and Organizational Psychology
The Pennsylvania State University
University Park, PA
Email: sth11@psu.edu

**Scarlett Miller**
School of Engineering Design, Technology and Professional Programs
The Pennsylvania State University
University Park, PA
Email: shm13@psu.edu

## ABSTRACT

*In this paper, we propose a new design variety metric based on the Herfindahl index. We also propose a practical procedure for comparing variety metrics via the construction of ground truth datasets from pairwise comparisons by experts. Using two new datasets, we show that this new variety measure aligns with human ratings more than some existing and commonly used tree-based metrics. This metric also has three main advantages over existing metrics: a) It is a super-modular function, which enables us to optimize design variety using a polynomial time greedy algorithm. b) The parametric nature of this metric allows us to fit the metric to better represent variety for new domains. c) It has higher sensitivity in distinguishing between variety of sets of randomly selected designs than existing methods. Overall, our results shed light on some qualities that good design variety metrics should possess and the non-trivial challenges associated with collecting the data needed to measure those qualities.*

*Address all correspondence to this author.

## NOMENCLATURE

SVS Design variety metric proposed in Shah *et al.* [1]
NM Design variety metric proposed by Nelson *et al.* [2]
HHI Herfindahl–Hirschman Index [3]

## KEYWORDS

Variety, Design Metrics, Optimization, Herfindahl index, Submodularity, Diversity, Crowdsourcing

## INTRODUCTION

Creativity is the capacity to generate unique and original work that is useful [4–6]. Creativity is useful at both individual and societal levels. At the individual level, creativity helps in effectively solving day-to-day tasks. At a societal level, it can yield meaningful scientific findings [5]. A well-known outlook relates creativity with divergent thinking — the capacity to produce a wider variety of ideas with higher fluency. Divergent thinking has been shown to correlate with the success of the final product [7–10]. Prior work supports that chances of solving a

problem increase when a more diverse set of ideas is produced in the initial stages of the design process [1, 11, 12]. These findings encourage the need to explore the design space in the early stages of design [13]. But how does one quantify design space exploration?

Engineering researchers have sought to capture how "explored the solution space" is by measuring design variety (pg. 117, [1]). There are two approaches typically deployed in engineering literature to measure design variety: subjective ratings of variety and a genealogical tree approach. As one example of subjectively evaluating design variety, Linsey *et al.* [14] proposed taking a set of ideas and dividing them into pools based on intuitive categories created by the coder. The metric relies on a rater's mental model rather than a quantitative procedure [1]. While these subjective ratings provide a relatively efficient method for measuring design variety in terms of the amount of time and effort required to code design variety, this efficiency comes at the potential cost of the validity and reliability of the metric.

In contrast to subjective ratings, the other approach to measure design variety is using a genealogical tree approach. In these approaches, subjective human raters are replaced with a deterministic formula that depends on the attributes of a set of designs. One of the first metrics to use this approach was developed by Shah, Smith and, Vargas-Hernandez [1] (SVS metric) who broke each design into four hierarchical levels (physical principle, working principle, embodiment, and detail) to calculate design variety. The SVS metric is repeatable and attempts to reduce subjectivity by using predefined criteria for measuring variety. However, many researchers have criticized it due to its lack of sensitivity and accuracy. For example, the genealogical tree calculation method (like SVS) has been shown to be inconsistent with experts ratings of variety [15]. In addition, studies have shown that the sensitivity of the SVS metric diminishes when it is applied to large datasets [16] due to the exclusion of important abstract differences and generally focuses on dissimilarity in the embodiment level [17].

In this paper, we reexamine these hierarchical metrics and compare them to methods of calculating diversity from other (non-engineering) domains. Specifically, we compare the tree-based measures of SVS [1] and NM [2] with the Herfindahl–Hirschman Index (HHI), which is a statistical measure of concentration [3, 18]. The HHI accounts for the number of firms in a market, as well as their concentration, by incorporating the relative size (that is, market share) of all firms in a market. HHI is used by the Department of Justice and the Federal Reserve in the analysis of competitive effects of mergers. For a market with $N$ firms, HHI is calculated by squaring the market share ($MS_i$) of all firms ($i \in \{1, \cdots, N\}$) in a market and then summing the squares, as follows:

$$HHI = \sum_{i=1}^{N} (MS_i)^2 \tag{1}$$

The key idea behind this metric is that market with more concentration will have a few large square terms. By comparing HHI to SVS [1] and NM [2], this paper argues and empirically demonstrates that HHI is a more accurate measure for variety that has clear benefits for engineering and design measurement applications. Specifically, the key contributions of this paper are:

1. We propose a new variety metric based on the Herfindahl–Hirschman Index and show that it better aligns with human judgments of variety compared to [1] and [2].
2. The metric function is monotone non-decreasing and supermodular, which allows us to propose a scalable greedy optimization algorithm with a constant factor guarantee of optimality. The greedy algorithm makes locally optimal choice at each step and guarantees that the final solution's variety will be atleast 0.63 of the highest variety solution. This allows us to find sets of ideas with high variety from a large collection in polynomial time.
3. We show that SVS and NM metrics give the same variety score to a large percentage of sets, while HHI index has higher sensitivity in distinguishing between different sets of ideas.

## BACKGROUND AND RELATED WORK

Before diving into the specific variety metrics we evaluate in this paper, we will first review the overall mathematical qualities that a good variety metric should possess. Next, we will discuss what factors should be considered in constructing a ground truth evaluation set for judging between different variety metrics. Lastly, we will end the section by reviewing past work on design variety metrics.

### Qualities of a good metric

Quality control is essential when creating and evaluating metrics that map abstract concepts like creativity to quantitative measures. Particularly when metrics can be either subjective and objective in scientific research, we need to demonstrate both the reliability and validity of such metrics without circularity [19], as well as reduce subjectivity in measurement techniques. For example, in the field of psychometrics, researchers try to craft sets of questions that produce internally consistent results — that is, if one asks the same questions one should get repeatable, similar answers even under minor changes to the test environment or experimental setup [20]. However, this only implies repeatability and not validity. Validity refers to the extent to which a

measurement reflects the absolute state of an artifact under observation — the ground truth). The term "valid" implies an external frame of reference or a universally accepted standard against which a measurement is tested [21]. There is a wide range of creativity metrics that leverage a rater's expertise in a given domain to ensure metric validity. This is necessary to eliminate circularity or measuring unvalidated metrics against other unvalidated metrics [22].

The key assumption in the past research is that raters who have considerable experience in a given domain are best suited to provide a ground truth for tasks like evaluating creativity. We obtain this ground truth from real world human evaluations, which can be used to measure the accuracy of any new metric. However, only using experts is no panacea. Expert time and effort is a scarce commodity, and this forces researchers to develop objective metrics that can aid quasi-experts or novice raters in accurately evaluating processes and ideas. The central hypothesis of that past work (which this paper also shares) is that by validating objective metrics against expert raters, the joint predictive power of expertise and repeatable objective research methods will outperform either by themselves.

Metrics used to measure variety like SVS and NM, aim to reduce subjectivity on the rater's part, to increase robustness in the processes used to analyze designs. When a metric is created, it is important to establish some desiderata (qualities we want) and acceptable qualities the metric must possess to ensure we obtain reliable results upon its execution. One example of establishing acceptable qualities of a metric was the work of Simonton and Amabile [23], who were key in standardizing the measurement of creativity in psychological research. Previously, most methods utilized pencil and paper tests, personality tests, biographical inventories (such as Schaefer and Anastasi's biographical inventory [24] and Taylor's Alpha Biographical Inventory [25]) and behavioral tests such as Torrance Tests of Creative Thinking. These tests were debatable in experiments that sought to reduce within-group variability and generally lacked a clear definition of creativity and an effective strategy to avoid biases on behalf of the rater [23].

Good metrics are required to have the ability to establish ground truths using expert agreements and must be replicable by other raters who use the metric. For subjective metrics, high inter-rater reliability and internal consistency are some of the desired qualities of the metric [26]. We argue that for any new variety metric, repeatability, validity, and explainability are also desirable qualities. If ground truth estimates of a quantity are available, then a new metric should align with this ground truth and the measurements should be repeatable. Variety metrics should also give explainable scores, that is, it should be possible to explain why one set of designs received a higher score than another set using a given metric.

## Design Variety and its Importance

The measure of design variety in engineering was introduced as a means to measure how well someone explores the solution during a design task [27]. The measure of design variety is important because research has shown that "there is no way to generate an optimum solution without exploring the solution space through early tentative ideas" (Pg.11 [28]). Generating a large number of ideas with iterative or small changes does not result in effective concept generation or innovative products. Hence, the potential to develop ideas of broad variety is correlated with the ability to successfully reconstruct and solve problems. This ability is referred to as cognitive restructuring in psychology [1] which has been used to counterbalance the number of ideas developed (quantity) in engineering design research because increases in the fluency of ideas must also be proportional to increases in the spread of the ideas [7].

Without exploration, designers may misconstrue the solution space to be very narrow [27]. One of the main contributing factors to this trend is functional fixation, or a blind adherence to solutions that are familiar and comfortable, which can generally lead to products of lower quality or innovation [29, 30]. As such, it is not surprising that research in engineering design has shown a correlation between the amount of design space explored and the quality of the final design [9].

Measurement of design space explored requires measuring mathematical functions on groups of ideas [31]. To address the desire to measure the extent to which tools promote variety, [1] developed a metric with the intent to provide a repeatable and reliable method to calculate design variety by rewarding ideas that are differentiated at higher levels of abstraction. In SVS metric, the authors decompose design variety into four hierarchical levels: the physical principle, followed by the working principle, embodiment, and detail. Specifically, they proposed that design variety should be calculated as shown below in equation 2.

$$V = \sum_{j=1}^{m} (f_j) \sum_{k=1}^{4} (S_k.B_k)/N \tag{2}$$

where $V$ is the variety score, $m$ is the number of functions solved by the design, $f_j$ is a weight assigned to the relative importance of function $j$, $S_k$ is the score for hierarchical level k, $B_k$ is the number of branches at hierarchical level $k$, and $N$ is the total number of ideas in the set. The key intuition behind this metric is that each idea is represented by hierarchical functions or attributes. Attributes on top of the hierarchy are more important than ones below, and if a set has multiple ideas with unique higher level attributes, then that set gets a higher variety score.

SVS metric has been criticized for double counting ideas at each level in the tree and for the selection of the weights at each level of the tree [2, 32]. Because of these pitfalls, Nelson

*et al.* [2] refined the metric by seeking to account for the double counting of ideas present in the SVS metric by considering the number of differentiation at each hierarchical level rather than considering all the levels. In addition, Nelson *et al.*modified the SVS metric by altering the weighting scheme from 10, 6, 3 & 1 to 10, 5, 2 & 1 for the physical principle, the working principle, the embodiment, and detail respectively. They argued that the new weighting scheme assures that at least two ideas at a lower hierarchical level must be added to equal the variety gain by adding a single idea at the next higher hierarchical level [2]. However, both SVS and NM do not provide a definition for each level of the hierarchy. There have been insufficient justifications for weights used in genealogical tree metrics [15] which can lead to large variations in the deployment of the metric in engineering design research. Other improvements of SVS metric includes the work of Verhaegen *et al.* [33], who combined Shahs metric with a Herfindahl index based tree entropy penalty, to encourage "uniformness of distribution" — essentially preferring trees that have even branching. Outside design, researchers have measured the breadth of ideation using metrics like mean pairwise distance between ideas [34] or by manually subgrouping functions into categories [35].

The new metric proposed in our work is closest in scope to Fuge *et al.* [36], who showed that both SVS and Verhaegen's metric were instances of submodular functions and argued that variety metrics are coverage functions which should belong to this family of functions. They introduced a probabilistic model that computes a family of repeatable variety metrics trained on expert data. In this work, we propose a new metric based on the Herfindahl index, which does not necessitate finding hierarchical features. Our metric also satisfies the properties of supermodularity (function whose negative is a submodular function), which allows us to optimize variety using a greedy heuristic algorithm. We show that unlike past metrics, this new metric has better alignment with judgment of variety by people.

## METHODOLOGY

In this section, we first describe a variety measurement method using the Herfindahl–Hirschman Index. Next, we show an example calculation of variety using the new metric. We show that the new metric can be optimized using a simple greedy algorithm to find sets of ideas with the highest variety.

### The Herfindahl–Hirschman Index for Variety

Over the last twenty years, economists have become increasingly interested whether diversity among multiple distinct population groups enhances or impedes a society's economic and social development. To quantify the economic impact of diversity, one must first create a proper index that captures how one society divides into various factions or parts.

Starting from the Gini index [37], Economists have used various diversity indices to evaluate the degree of social, economic, cultural, and other dissimilarities among people, regions, and countries. Initially used as an income inequality measure, the Gini index was re-interpreted by Simpson [38] as the inverse Hirschman–Herfindahl index. That index measured industry concentration and was also used by Greenberg [39] for the measurement of linguistic diversity. The value of the index measures the probability that two randomly chosen individuals in society belong to different groups.

This Herfindahl index (also known as Herfindahl–Hirschman Index, HHI, or sometimes HHI-score) measures a firm's size relative to the industry and indicates the amount of competition among firms. We described the mathematical structure of HHI in Eqn. 1 above. In this section, we propose a variant of HHI-score that can measure the variety of a set of designs.

To do so, we assume that we are given a set of designs $S$. Each design within set $S$ is divided into hierarchical levels like functional principle, working principle, embodiment, and detail (similarly to SVS and NM above). We then calculate the HHI index for each level separately for the entire set. For example, the HHI index for the functional principle level is given by:

$$HHI_F(S) = \frac{\sum_{i=1}^{N_f} |FP_i|^2}{N^2} \tag{3}$$

In this, $|FP_i|$ is the number of designs using functional principle $i$ and $N_f$ is the total number of functional principles. $N$ is the total number of designs in the set $S$. $HHI_F(S)$ varies between $1/N$ to 1. Similarly, we can define HHI for working principle, embodiment and details. Hence, the total HHI variety metric is a weighted sum of these four metrics as follows:

$$\begin{aligned} HHI(S) = w_1 \frac{\sum_{i=1}^{N_f} |FP_i|^2}{N^2} + w_2 \frac{\sum_{j=1}^{N_w} |WP_j|^2}{N^2} + \\ w_3 \frac{\sum_{k=1}^{N_e} |EM_k|^2}{N^2} + w_4 \frac{\sum_{l=1}^{N_d} |DE_l|^2}{N^2} \end{aligned} \tag{4}$$

Here, $HHI(S)$ is the total HHI score for a set of designs $S$. The weights $w_1$, $w_2$, $w_3$ and $w_4$ can be chosen such that the resultant value is between 0 and 1 (sum of weights is 1). For instance, if all factors are equally important, then $w_1 = w_2 = w_3 = w_4 = 1/4$. $|WP_j|$ is the number of designs in the set using working principle $j$, $|EM_k|$ is the number of designs using embodiment $k$ and $|DE_l|$ is the number of designs using detail level $l$. $N_f$, $N_w$, $N_e$ and $N_d$ are the total number of functional, working, embodiment and detail principles. Note that the normalized definition of HHI using proportions is not

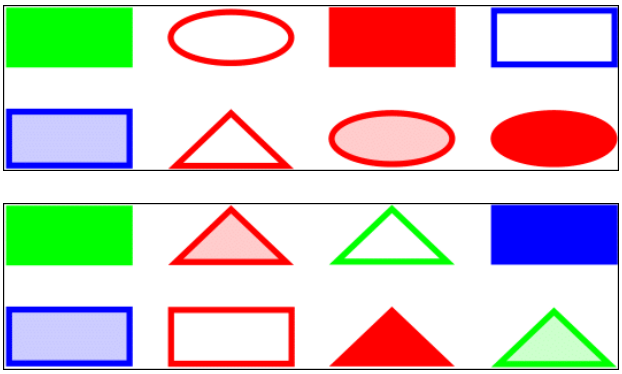

**FIGURE 1**. Example of two polygon sets (Top shows Set A and bottom shows Set B) shown to participants in our experiment. Participant answers the question: "Which set is more diverse?"

a supermodular function [1]. HHI metric defined by us in Equation 1 is supermodular when it is not normalized by $N^2$, which we use optimization. This means that when a design is added to a larger set, the increase in HHI score is larger compared to the case when the same design is added to a smaller set. This property can be exploited to find sets of maximum diversity using a greedy algorithm [40], which guarantees that the variety of the greedy search solution will be within 63.2% (or $1-\frac{1}{e}$) of the variety of the optimal solution.

### Calculating variety of a set

To demonstrate the calculation of HHI, we take a set of designs shown in Fig. 1 as an illustrative toy example.

In Fig. 1, for the set shown on top, there are eight polygons ($N = 8$). There are four items with a rectangular shape, three items with an oval shape and one triangular. There are five red colored polygons, two blue and one green. Three items have a solid fill, two have shaded and three are empty inside. Without loss of generality, for this example, we assume that color is the functional principle of a polygon, shape is the working principle and shading is the embodiment. We assume that all three levels are equally important in deciding the variety of Set A ($w_1 = w_2 = w_3$) and $N_f = 3$ as there are three unique functional principles (color). The $HHI_F$ score for color will be $(5/8)^2+(2/8)^2+(1/8)^2 = 0.47$. Similarly, $HHI_W$ score for shape will be $(4/8)^2+(3/8)^2+(1/8)^2 = 0.39$ and $HHI_E$ score for fill will be $(3/8)^2+(2/8)^2+(3/8)^2 = 0.34$. As all features are equally important, then HHI for the set of designs will be the average of the three numbers $(0.47+0.39+0.34)/3 = 0.40$. Similarly, the variety of any set of designs can be calculated.

[1] Submodular functions are functions defined over sets that are designed to model diminishing marginal utility, which is the mathematical property one needs to model diversity or variety [36]. Supermodular functions are functions whose negative is a submodular function.

### Optimizing variety of a set

Using metrics like SVS, NM and HHI we can measure the variety of a given set of ideas (like the sets shown in Fig. 1). However, what happens when we want to choose the set of eight polygons which have the maximum variety? One way is to enumerate all possible sets of size eight (about 2.2 million sets), calculate their variety score and then find the set with highest variety. This approach becomes intractable as the number of items in the ground set increases.

Another approach, and the one we use, is to leverage mathematical properties of the variety function and find approximate solutions close to optimal. To find sets of maximum variety, we use a sub-modular greedy algorithm (Algorithm 1) to order the ideas [40]. Given the set $V$ of all ideas, the algorithm starts with an empty set $S = \{\}$ and add ideas to this set according to Algorithm 1. In the end, this set $S$ will be the ranking that the algorithm outputs. It will contain all ideas ordered in such a way as to maximize the objective value defined in Eq. 4 (when the function is not normalized by $N^2$), *i.e.*, the ideas of high variety (*i.e.*, from principles less represented so far) are at the top of the ranking.

To achieve this, the algorithm starts adding ideas to an empty set $S$ and removing them from set $V$, one idea at a time, such that the selected idea $i \in V$ is the one with the lowest marginal gain $\delta f(S \cup i)$ on set $S$. Here $\delta f(S \cup i) = HHI(S \cup i) - HHI(S)$. Here the set $V$ is the set of all designs and set $S$ is the selected set of design which we find using a greedy algorithm.

By choosing at each step to add the idea that will maximize variety (minimize the metric function) of the existing set of already added ideas, the algorithm not only selects the ideas but also orders them as well. Finally, as the function in Eq. 4 is super-modular and monotonic, the algorithm is also theoretically guaranteed to provide the best possible $(1-\frac{1}{e})$ polynomial-time approximation to the optimal solution [41, 42].

**Algorithm 1:** Greedy algorithm for maximum variety. The algorithm performs a polynomial-time greedy maximization of the gain on the non-normalized HHI variety index. The output is a ranking of all ideas such that high-variety ideas are at the top.

**Data:** Original set $V$ of all ideas
**Result:** Ranked set $S$ of all ideas
initialization;
$S \leftarrow \emptyset$;
**while** $V \neq \emptyset$ **do**
  Pick an item $V_i$ that maximizes $\delta f(S \cup i)$;
  $S = S \cup \{V_i\}$;
  $V = V - V_i$;
**return** $S$;

## EXPERIMENTS AND RESULTS

We conducted two experiments to benchmark the proposed HHI metric with the commonly used SVS and NM metrics: (1) an experiment using a known and easily verifiable ground truth based on polygons, and (2) an experiment using actual design sketches provided by engineering students and rated by domain experts. Before introducing these experiments and their main results and implications, we describe how we constructed our experimental dataset of set comparisons for these two domains. As we have shown, constructing such sets is non-trivial, and one contribution of this paper lies in describing a procedure for constructing such comparison sets for new domains.

### Estimating Design Variety Ground Truth using Human Pairwise Comparisons

The first step in vetting design rating metrics is to identify a 'ground truth' of the measure that the metric is trying to capture and then calculate how accurate any given metric is in capturing that ground truth. However, for the case study presented here (design variety), ground truth estimation is difficult due to the large combinatorial space for sets of items and the lack of a benchmark dataset. For instance, a small set of thirty design ideas has more than one billion possible sets of designs for which variety can be calculated. Exhaustive calculation of ground truth is impossible. Secondly, we do not use any existing variety metric to create the ground truth. Doing so would make the assumption that a given metric represents true variety, which is what the ground truth is used to establish. Instead, we propose the development of a ground truth using pairwise human judgments.

To establish a ground truth dataset for the calculation of design variety, we first need three components:

1. A ground set of design items over which sets are created
2. Sets of designs derived from the ground set for which variety scores are calculated
3. Tree annotations for each design item so we can calculate tree-based metrics

Variety scores are calculated on a set of designs. However, human raters are not good at giving absolute scores [43] due to differences between internal scales of subjects, a well-known problem for subjective pairwise scaling. For instance, given the set of designs shown in Fig. 2, it would be difficult for a human rater to say whether this set of six designs scores 6 out of 10 or 8 out of 10 for variety. Different raters may also use different internal scales.

In contrast, if we ask a rater to rate whether they find the variety of set shown in Fig. 2 Set A greater than the variety of those shown in Fig. 2 Set B, they may answer it relatively easily. Hence, we propose that ground truth for variety should be created using pairwise queries, where each query contains two sets and one set is voted by human raters to have higher variety compared to the other set. To elicit responses from experts, we give them two sets at a time and ask them for pairwise comparisons of the form: "Which set of designs has higher variety?"

### Measuring Variety for Polygons

In this experiment, we compared the performance of HHI, SVS and NM metrics in measuring the variety of a set of polygons. We first create a base set of 27 polygons. Each polygon has three attributes — shape, color, and shading. Each attribute can take three unique values. Polygons can be rectangular, triangular or oval shaped. They can be red, blue or green colored. Shading varies between polygons as complete fill, shaded or empty.

The number of possible sets of polygons is very large ($2^{27}$), hence calculating the variety score of all possible sets is not feasible. Instead, we narrow down our search to focus on three set sizes: when the number of items in a set is 4, 6 and 8. If we ask human raters to compare sets with larger than eight items, the task becomes very difficult for them. For a given set size, we first randomly pick 100 sets for comparison. From these 100 sets, we calculate all possible pairwise comparisons (4950 comparisons). Next, we calculate SVS, NM, and HHI scores for each set. For SVS and NM, we assume that 'Color' is the functional principle, 'Shape' is the working principle and 'Shading' is the embodiment.

**Result 1: Existing metrics cannot distinguish between sets.** Table 1 shows the percentage of comparisons where each metric finds both the sets of equal variety. We note that SVS and NM metrics do not distinguish between a large percentage of comparisons (up to 37%), while HHI gives identical scores to a much smaller percentage of pairwise comparisons. This implies that existing metrics are not sensitive or discriminative to differences between sets.

**Result 2: Existing metrics vote similarly to one another.** Table 1 also shows the percentage agreement between different metrics. We see that SVS and NM vote similarly for 80-85% of set comparisons for various set sizes. This means that for a large proportion of comparisons, both metrics are indistinguishable as they give the same pairwise response. If SVS finds Set A has higher variety, then so does NM. In contrast, the agreement between HHI and other metrics is close to random. Due to the lack of benchmark dataset, it is difficult to comment on whether a lack of agreement between metrics is a good thing or not. We show later in the results, HHI aligns with the human raters more than SVS and NM for two datasets.

To establish a ground truth for comparing different metrics, we proceeded with the following steps. First, we selected pairwise comparisons where SVS and NM could actually distinguish between the two sets; that is, both the metrics did not calculate the same variety score to both sets. This is important since we

| | Same Score | | | Agreement | | |
|---|---|---|---|---|---|---|
| Method | SVS | NM | HHI | SVS-NM | HHI-SVS | HHI-NM |
| Size 4 | 27.3% | 37.0% | 15.8% | 84.4% | 54.2% | 50.2% |
| Size 6 | 31.7% | 21.4% | 14.7% | 81.0% | 47.6% | 50.0% |
| Size 8 | 28.5% | 12.9% | 10.9% | 82.5% | 49.4% | 56.9% |
| Size 10 | 31.2% | 14.5% | 9.2% | 84.4% | 54.2% | 50.2% |

**TABLE 1**. a) Percentage of pairwise comparisons when design metrics give same score to both designs. Lower percentages are good as it indicates that a metric can distinguish between sets. We notice that SVS metric gives same score for approximately 30% of the sets. b) The right side shows agreement between metrics for pairwise comparisons. We notice that SVS and NM tend to vote similarly for more than 80% of the sets.

want any collected human judgment to differentiate existing metrics, and we cannot do this if we select comparisons where the two metrics calculate the same value. Secondly, we select the sets where both metrics disagreed on their vote. This means if SVS voted Set A to be higher variety, then NM would vote Set B to be higher variety. Note that this is a small set of pairwise comparisons — as we noted from Table 1, both metrics vote similarly for more than 80% of the comparisons.

Finding human annotations for such sets allows us to find out which of the two metrics better aligns with human responses. Finally, we take the top 5 sets where SVS is most confident that one set has higher variety than another and the top 5 sets where NM is most confident that one set has higher variety than another set (*i.e.*, the difference between the scores are maximum). We combine these two to generate 10 queries which are then given to human raters.

To find the ground truth for polygons, we conducted an Amazon Turk study, in which we collected responses from crowd workers for pairwise queries. A sample query with two sets of eight polygons is shown in Fig. 1. Judging the variety of polygons does not require expertise in the area and Amazon Turk allows us to gain a large number of responses. We collected pairwise responses for three different set sizes. For each set size, we created ten pairwise queries. For each query, we collected ten responses from Amazon Turk participants. We randomized the order of the queries and also the order of the options shown to different participants to reduce the possibility of any ordering bias. We subdivided the surveys into two parts to reduce fatigue. No worker was repeated across surveys and six queries were repeated to filter out workers with very low internal consistency.

**Result 3: Human raters largely agree on what it means to have a high variety set of polygons.** The survey responses showed that on average people had consensus on one set being more diverse or higher variety than another set. The number of votes received by the set pairwise query receiving a majority vote for sets of size 4 was:[9, 8, 9, 7, 6, 9, 8, 6, 8, 7] respectively. This means that for the first query, 9 people out of 10 voted for the same set. For the second query where two sets of size 4 were shown, 8 people voted for the same set as being of higher variety. Similarly, for sets of size 6, [5, 5, 9, 9, 9, 8, 6, 8, 5, 8] votes were received by the majority set and [7, 5, 7, 7, 9, 9, 8, 6, 7, 6] votes were received by the majority set for sets of size 8.

A direct comparison between SVS, NM, and HHI metrics using the published weights would be unfair to SVS and NM, as HHI weight parameters can be optimized specifically for each domain. The published weights for SVS metric is [10, 6, 3, 1] and published weights for NM metric is [10, 5, 2, 1]. Hence, we give the same flexibility to SVS and NM metrics by allowing the weights of functional principle, working principle and embodiment to be optimized to maximize their performance. For a given metric (say SVS) and weight combination (say 4, 3, 3), we calculate the variety scores for both sets in a given pairwise comparison. Suppose we had total 10 humans who voted on a pairwise comparison. If SVS metric finds that Set A has more variety than Set B, and 8 humans had also voted this way, we allocate all these votes to SVS metric. If the metric found Set B has higher variety than Set A, then this metric receives the 2 votes which humans gave to the other set. As we ask 30 different queries from people, to judge the metric, we aggregate votes for all 30 queries.

For our experiment, the maximum number of votes that any metric can receive is 220 — that is if it always votes with the majority opinion of human raters. Note that in an ideal world, if all humans voted for the same set for all 30 queries, the maximum number of votes that any metric can receive would be 300. Suppose a metric receives 200 votes in total, then we say that it has 90.9% alignment (100x200/220 = 90.9) with human ratings.

**Result 4: HHI outperforms SVS and NM w.r.t. human agreement on polygon variety.** Table 2 shows the comparison between SVS, NM, and HHI for alignment with human ratings. We find that SVS and HHI have similar best case performance. We varied the weights of each functional level between 1 to 10 in steps of 1, giving us 1000 possible performance scores corresponding to each weight combination [w1, w2, w3]. We find that HHI performs better than SVS in the median case. The median case is calculated over all thousand weight combinations.

From Table 2, we can conclude that HHI aligns with human perception of variety to the highest degree, irrespective of the choice of weights — that is, its performance is robust to weight choices. Even in the worst case, HHI aligns with 74.5% of human ratings. We find that the highest performance is obtained for many combinations of weights like 1, 2 and 10. SVS performs similarly, however, we generated these comparisons such that SVS has high confidence in its choice between both the sets (by design). In contrast, if we select sets to compare at random, SVS calculates the same score for more than one-fourth of the queries. This drastically reduces the SVS performance in align-

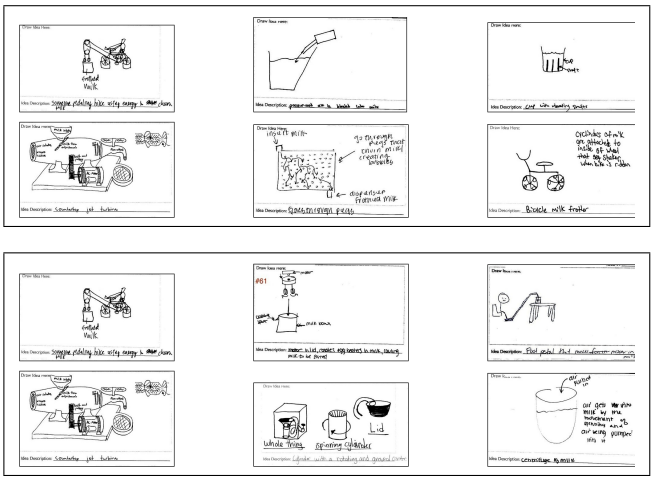

**FIGURE 2**. Top: Sample of Set A where all raters agreed it was more diverse than Set B. Bottom: Sample of Set B where all raters agreed it was less diverse than Set A.

ment with human responses — humans would have a clear preference between the variety of two sets, but SVS would be indifferent. Hence, the HHI metric outperforms both SVS and NM in alignment with human's judgment of variety.

| Method | Median Case | Best Case | Worst Case | Sample optimal weights |
|---|---|---|---|---|
| HHI | 81.8% | 95.4% | 74.5% | 1, 2, 10 |
| SVS | 79.0% | 95.4% | 59.0% | 2, 1, 1 |
| NM | 54.5% | 86.3% | 40.9% | 10, 3, 1 |

**TABLE 2**. Comparison of design variety metrics in alignment with human ratings

## Measuring Variety for Milk Frother Sketches

To measure the variety of milk frothers, we gathered data from a previous experiment conducted by Starkey, Hunter, and Miller [44], which consisted of 934 ideas. Specifically, the data set consisted of ideas developed by 89 first-year students from an undergraduate engineering course and 52 senior students from a capstone engineering course including 95 males and 46 females. The ideas developed in this dataset were from a design task where participants were asked to generate ideas for a "novel and efficient milk frother." This task was selected because the task addressed solving a product-based problem.

To create the dataset of sets of milk frother sketches, we used the ground set of ten design sketches studied in Ahmed *et al.* [45]. The benefit of using these ten sketches is the availability of tree annotations as well as information in the form of subjective idea maps, which we use when discussing the final results below. The total number of possible sets for these ten sketches is 1024. We first calculate the variety scores for all these sets using SVS and NM metric.

Similar to the polygon case, to create a ground truth dataset of pairwise queries, we first want to find the queries which are most meaningful. However, in this case, we also have information about Euclidean embeddings for each sketch as discussed in [45]. These embeddings are essentially 2-D maps with each design having x and y coordinates allocated to them. Similar designs occur closer to each other than dissimilar designs on this map. To find which sets to ask humans to rate, we use three metrics: SVS, NM and average pairwise distance of a set. The last metric is derived using an embedding of designs derived in the study by Ahmed *et al.* [45]. The design embedding was picked randomly (as each participant in the study had a different design embedding and we needed only one design embedding to guide our experiment) and it provides the 2-D positions for each sketch and is only used to guide the selection of sets to be shown to human judges. The choice of the design embedding does not alter the key findings of this section as it is only used to guide the selection of queries which are asked from people. Using these ten sketches, our goal is to create pairwise queries with sets of six sketches each. We decided to create the ground truth with pairs of six images as the median number of sketches made by a participant in our dataset was six. The number of sets of size 6 is 210 unique combinations. We calculated the variety scores for all combinations and rank ordered them from the highest variety set to the lowest variety set using the pairwise average distance metric.

Out of these 210 sets, we obtained 21,945 pairs of sets and calculated the absolute rank difference between the two items for each comparison. A small rank difference implies that the two sets have similar variety, while a large rank difference implies that the metric is confident that one of the set has a significantly higher variety than the other. After calculating the rank differences, we selected 20 comparisons based on two factors. First, we should select comparisons where each metric (pairwise distance, SVS, and NM) votes differently on which set has higher variety — *i.e.*, if all ratings agree on the comparison, then human expert ratings would not discriminate them. Second, we should select sets with a high-rank difference, but that also differ from sets we are using in other selected comparisons. That is, we want to ensure that a metric is confident in its vote, but that we also get good coverage over different types of sets in the data by ignoring pairs that have already been selected.

Among these candidate sets, we select 20 pairwise queries that are given to four expert raters using a Qualtrics survey. We repeat two comparisons (10% repeated queries) in each survey to measure the internal consistency of each expert, giving them a total of 22 queries. Experts can choose whether Set A is higher variety compared to Set B or they can select the option of 'Can't decide'. From these expert ratings, we find that all four experts agreed on 9 out of 20 queries, while at least three experts agreed

on 15 queries. Due to a majority agreement on these 15 queries, we select them as the ground truth dataset for comparing variety metrics. Next, we use this ground truth dataset to compare the SVS and NM metrics.

**Result 5: SVS and NM are equivalent to random chance, w.r.t. matching expert assessments of milk-frother variety.** We find that both SVS and NM align with only one-third (33.3%) of our human-provided ground truth dataset — that is five comparisons. We also change the weights for SVS and NM and report how close these metrics are to human experts. To explore the sensitivity of these results, we calculate the NM and SVS scores for every valid weight combination used by each metric. Using these weights, we find that SVS aligns with 33.3% of the pairwise expert assessments of milk-frother variety irrespective of the weights used — that is, changing the tree weights used by SVS has zero effect on whether or not it agrees with human experts. NM aligns with 33.3% of the dataset for 95.6% of all the weight combinations. For the rest, it has no alignment with any expert ratings — that is, NM's scores are more sensitive to its internal weights, but not in a way that benefits its score accuracy with respect to human raters. The alignment scores are close to random chance for three categories (Greater, Smaller and Equal) showing that SVS and NM are unable to capture human intuition of variety for the examples we tested.

**Result 6: HHI robustly outperforms SVS and NM w.r.t. human comparisons, but still has a non-trivial error.** In contrast to SVS and NM, HHI aligns with 9 out of 15 comparisons when weights are optimized for each level. We find that many weight configurations for HHI lead to highest performance, including w=[1, 9, 5].

Hence, HHI aligns with human judgment of variety more than both SVS and NM metrics for two standard datasets. However, it still is not 100% accurate with respect to human benchmarks. However, we had assumed that the annotations provided for SVS, NM, and HHI for different hierarchical levels are accurate. If this is not the case, any variety metric will have a large error as it may not capture the true features. The construction of hierarchical trees is outside the scope of this paper but it is important to understand that metrics may be limited by the specific choice of how one constructs a tree, which also needs to be verified.

We propose that by using our above method for constructing these ground truth variety comparisons, future papers will be able to use these and other ground truth variety pairwise comparisons to judge the comparative quality of other metrics as well. This would provide a common scale over which metrics are compared.

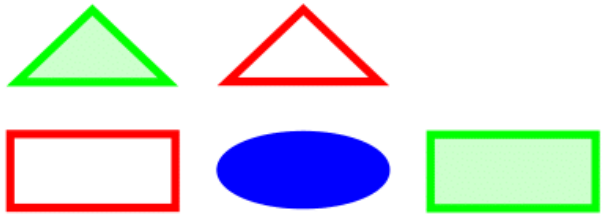

**FIGURE 3**. Set of five polygons with highest variety found using a greedy algorithm applied to supermodular objective function

### Finding Sets of Designs with Highest Variety

One of the auxiliary outcomes of using a HHI derived index for variety measurement is that it provides a simple method to find the highest variety sets. Suppose you want to find a set of five polygons which have the highest variety from a given set of 27 polygons. Using existing NM and SVS metrics, the only way to do so is to enumerate all 80730 (27 choose 5) possible sets of five polygons, then calculate their NM and SVS scores and find the set with the highest score. This approach becomes infeasible when the ground set becomes large (for example 2.5 Billion sets for 200 designs) due to a large number of possible options (mathematically, this is because the problem is NP-Hard).

In contrast, we use Algorithm 2 to rank order all polygons or to select a subset. The resultant set is shown in Fig. 3. The set has high variety with respect to color, shape, and shading. The method selects one polygon at a time based on which polygon provides lowest marginal gain. As mentioned above, this is possible in polynomial time due to the supermodular behavior of HHI.

## DISCUSSION

Our experiments highlight several broader implications, both around how variety metrics are constructed and verified, as well as in how existing metrics are used across domains.

### Selecting appropriate validation sets for variety measures is non-trivial

As we showed above, selecting exactly which sets of designs to show experts for ground truth labeling is non-trivial. First, the combinatorial nature of the problem (sets of designs) makes exhaustive labeling by experts impractical for anything above a handful of designs. But randomly sub-sampling this combinatorial set does not solve the problem: many metrics may trivially agree on a large portion of the space.

We proposed possible desiderata on what comparisons to show experts, as well as several potential methods to make this selection, such as maximal rank order disagreement, distances over embedded spaces computed via past techniques [45], and

space coverage over different sets. Constructing comparisons in this fashion does lead to potential bias: as we saw in Result 4, by preferentially sampling sets where metrics were confident in their answers, we may, in fact, overestimate their performance with respect to their average performance in practice.

The trade-off here is one of time and cost. If one picks comparisons to maximize discriminative power among metrics, this will inevitably ignore portions of the space where they agree and inflate performance measures. In contrast, if one does not do this one may collect many expensive expert comparisons that, while covering the space well, do not provide much value in separating good metrics from bad ones.

One limitation of our proposed approaches is that we currently provide no theoretical guarantees regarding the number or scope of queries needed to achieve a certain assessment accuracy. The number of comparisons we collected above was driven by primarily practical concerns — how many expert comparisons could we realistically expect to collect in our available time budget? Future work could address how to perform this collection in an optimal fashion (*e.g.*, using Active Learning) and to bound the number of comparisons one would need to collect.

### Good variety metrics need to be accurate and discriminative

As we showed in Results 1 and 2, good metrics need to not only be accurate but also highly discriminative or sensitive. We found that commonly used metrics can lack sensitivity across a broad range of comparisons. Even if such metrics are accurate, they have limited usefulness as measurement instruments — that is, they cannot detect small effect sizes in terms of differences in variety. We argue that, in addition to focusing on accuracy, future metric development should compute and account for the sensitivity of the measurement instrument for the given domain, and such quantities should be reported in subsequent papers.

### Metric performance can differ significantly across domains

Comparing Results 4 and 5, we see that a given metric applied to one domain/problem may have drastically different performance. In our case, SVS performed well with respect to human comparisons on the polygon case, but poorly on the milk frother case. While it is perhaps obvious that a metric's accuracy depends on where it is applied, we note that, in practice, past researchers have broadly used existing metrics (both SVS, NM, and others) with limited to no verification and calibration of the measurement instrument to that domain.

We believe that our results here should give other researchers pause before blindly applying an existing variety metric to a new problem without first conducting some of the pairwise verification we detail above. We are releasing both the datasets we collected in this paper and the tools we used to construct human comparisons in the hope that future researchers will have an easier time constructing verification tests for new metrics or domains.[2] We believe that the proposed metric can be used in combination with other design metrics to provide insights from different perspectives of a set of designs. The usage of this metric and creation of new ground truth datasets should take into account the context that designers have deep knowledge in a field and can judge variety through different lenses and with an experience that may not always be possible from a quantitative metric.

### HHI is a promising alternative metric that allows optimization of variety

We demonstrated via Results 4 and 6 that using HHI matched or exceed the performance of commonly used metrics. This was true in both the Polygon and Milk Frother experiments. Calculating the HHI is computationally simpler to the benchmark tree-based constructions of SVS and NM.

More importantly, the supermodular form of HHI allows us to efficiently (*i.e.*, in polynomial time) approximate the highest variety sets of designs, given a corpus. For design corpora larger than approximately 50 designs, this leads to order-of-magnitude reductions in computational effort in finding optimal variety subsets of design, compared to existing metrics. The fact that HHI can be easily optimized to match human judgments for a domain makes it flexible to apply to different problems if one gathers pairwise comparison data as described above.

Future work could cast the fitting of HHI as an active learning problem to reduce the number of expert comparisons needed to fit HHI to a given domain.

## CONCLUSION

In this paper, we contributed: (1) a new design variety metric based on the Herfindahl index; (2) a practical procedure for comparing variety metrics via the construction of ground truth datasets from pairwise comparisons by experts; and (3) an empirical demonstration of this procedure and metric on two new two ground truth datasets using milk frother design sketches and polygons. Using this dataset, we then compared the performance of two existing and commonly used tree-based metrics and showed that our newly proposed metric aligns with human ratings more than existing metrics. As an ancillary benefit, we also show that by using a simple greedy algorithm our new metric can find sets of designs with the highest variety in polynomial time.

Overall, our results shed light on some qualities that good design variety metrics should possess and the non-trivial challenges associated with collecting the data needed to measure those qualities. These results provide guidance on how and when various commonly used metrics may or may not be valid, as well

[2] https://github.com/IDEALLab/design-variety

as a concrete scientific process by which to gain further insight into when and where metrics apply.

We hope that the procedures we outline here can provide a catalyst for deeper discussion regarding how we measure and verify variety within engineering design. We encourage researchers to build upon and contribute to the datasets we have started collecting and distributing for these problems. Our hope is that by better understanding how to measure variety and ultimately optimize variety, we will be able to reliably and scalably support designers in improving their creativity and competitiveness.

## ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1728086. We acknowledge the effort of both the MTurk workers and expert raters who help us collect ratings on the two experiments.

## REFERENCES

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

[2] 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.

[3] Hirschman, A. O., 1964. "The paternity of an index". *The American economic review,* ***54***(5), pp. 761–762.

[4] Amabile, T. M., 1996. *Creativity in context: Update to the social psychology of creativity*. Hachette UK.

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

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

[7] Torrance, E. P., 1972. "Predictive validity of the torrance tests of creative thinking". *The Journal of creative behavior,* ***6***(4), pp. 236–262.

[8] Acar, S., and Runco, M. A., 2017. "Latency predicts category switch in divergent thinking.". *Psychology of Aesthetics, Creativity, and the Arts,* ***11***(1), p. 43.

[9] Dylla, N., 1991. "Thinking methods and procedures in mechanical design". *Mechanical design, technical university of Munich, PhD*.

[10] Beitz, W., and Pahl, G., 1996. "Engineering design: a systematic approach". *MRS BULLETIN,* ***71***.

[11] Pahl, G., and Beitz, W., 2013. *Engineering design: a systematic approach*. Springer Science & Business Media.

[12] Dorst, K., and Cross, N., 2001. "Creativity in the design process: co-evolution of problem–solution". *Design studies,* ***22***(5), pp. 425–437.

[13] Henderson, D., Helm, K., Jablokow, K., McKilligan, S., Daly, S., and Silk, E., 2017. "A comparison of variety metrics in engineering design". In ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. V007T06A004–V007T06A004.

[14] Linsey, J. S., Clauss, E., Kurtoglu, T., Murphy, J., Wood, K., and Markman, A., 2011. "An experimental study of group idea generation techniques: understanding the roles of idea representation and viewing methods". *Journal of Mechanical Design,* ***133***(3), p. 031008.

[15] Linsey, J. S., 2007. "Design-by-analogy and representation in innovative engineering concept generation". PhD thesis, University of Texas, Austin.

[16] 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.

[17] 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.

[18] Rhoades, S. A., 1993. "The herfindahl-hirschman index". *Fed. Res. Bull.,* ***79***, p. 188.

[19] Shatz, D., 2004. *Peer review: A critical inquiry*. Rowman & Littlefield.

[20] Kline, P., 2014. *The new psychometrics: science, psychology and measurement*. Routledge.

[21] Twomey, M., Wallis, L. A., and Myers, J. E., 2007. "Limitations in validating emergency department triage scales". *Emergency Medicine Journal,* ***24***(7), pp. 477–479.

[22] Harnad, S., 2008. "Validating research performance metrics against peer rankings". *Ethics in science and environmental politics,* ***8***(1), pp. 103–107.

[23] Amabile, T. M., and Pillemer, J., 2012. "Perspectives on the social psychology of creativity". *The Journal of Creative Behavior,* ***46***(1), pp. 3–15.

[24] Schaefer, C. E., and Anastasi, A., 1968. "A biographical inventory for identifying creativity in adolescent boys.". *Journal of Applied Psychology,* ***52***(1p1), p. 42.

[25] Taylor, C., and Ellison, R., 1966. "Alpha biographical inventory". *Salt Lake City, UT: Institute for Behavioral Research in Creativity*.

[26] Cropley, A. J., 2000. "Defining and measuring creativity: Are creativity tests worth using?". *Roeper review,* ***23***(2), pp. 72–79.

[27] Dow, S., Fortuna, J., Schwartz, D., Altringer, B., Schwartz, D., and Klemmer, S., 2011. "Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results". In Proceedings of the SIGCHI Conference

on Human Factors in Computing Systems, Acm, pp. 2807–2816.
[28] Cross, N., and Roy, R., 1989. *Engineering design methods*, Vol. 4. Wiley New York.
[29] Jansson, D. G., and Smith, S. M., 1991. "Design fixation". *Design studies,* ***12***(1), pp. 3–11.
[30] Kershaw, T. C., and Ohlsson, S., 2004. "Multiple causes of difficulty in insight: the case of the nine-dot problem.". *Journal of experimental psychology: learning, memory, and cognition,* ***30***(1), p. 3.
[31] 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.
[32] Wilson, J. O., Rosen, D., Nelson, B. A., and Yen, J., 2010. "The effects of biological examples in idea generation". *Design Studies,* ***31***(2), pp. 169–186.
[33] Verhaegen, P.-A., Vandevenne, D., Peeters, J., and Duflou, J. R., 2013. "Refinements to the variety metric for idea evaluation". *Design Studies,* ***34***(2), pp. 243–263.
[34] Chan, J., Dang, S., and Dow, S. P., 2016. "Comparing different sensemaking approaches for large-scale ideation". In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, pp. 2717–2728.
[35] Chan, J., Fu, K., Schunn, C., Cagan, J., Wood, K., and Kotovsky, K., 2011. "On the benefits and pitfalls of analogies for innovative design: Ideation performance based on analogical distance, commonness, and modality of examples". *Journal of mechanical design,* ***133***(8), p. 081004.
[36] Fuge, M., Stroud, J., and Agogino, A., 2013. "Automatically inferring metrics for design creativity". *ASME Paper No. DETC2013-12620*.
[37] Gini, C., 1912. "Variabilità e mutabilità". *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*.
[38] Simpson, E. H., 1949. "Measurement of diversity". *Nature,* ***163***(4148), p. 688.
[39] Greenberg, J. H., 1956. "The measurement of linguistic diversity". *Language,* ***32***(1), pp. 109–115.
[40] Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L., 1978. "An analysis of approximations for maximizing submodular set functionsi". *Mathematical Programming,* ***14***(1), pp. 265–294.
[41] Feige, U., Mirrokni, V. S., and Vondrak, J., 2011. "Maximizing non-monotone submodular functions". *SIAM Journal on Computing,* ***40***(4), pp. 1133–1153.
[42] Krause, A., and Golovin, D., 2014. "Submodular function maximization". In *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, pp. 71–104.
[43] Kendall, M., 1962. *Rank correlation methods*. Theory and applications of rank order-statistics. Hafner Pub. Co.
[44] Starkey, E. M., Hunter, S. T., and Miller, S. R., 2019. "Are creativity and self-efficacy at odds? an exploration in variations of product dissection in engineering education". *Journal of Mechanical Design,* ***141***(1), p. 012001.
[45] Ahmed, F., Ramachandran, S. K., Fuge, M., Hunter, S., and Miller, S., 2019. "Interpreting idea maps: Pairwise comparisons reveal what makes ideas novel". *Journal of Mechanical Design,* ***141***(2), p. 021102.