Faez Ahmed[1]
Department of Mechanical Engineering,
Northwestern University,
Evanston, IL 60208
e-mail: faez@northwestern.edu

Sharath Kumar Ramachandran
School of Engineering Design,
Technology and Professional Programs,
The Pennsylvania State University,
University Park, PA 16802
e-mail: sharath@psu.edu

Mark Fuge
Department of Mechanical Engineering,
University of Maryland,
College Park, MD 20742
e-mail: fuge@umd.edu

Sam Hunter
Industrial and Organizational Psychology,
The Pennsylvania State University,
University Park, PA 16802
e-mail: sth11@psu.edu

Scarlett Miller
School of Engineering Design,
Technology and Professional Programs,
The Pennsylvania State University,
University Park, PA 16802
e-mail: shm13@psu.edu

Design Variety Measurement Using Sharma–Mittal Entropy

Design variety metrics measure how much a design space is explored. This article proposes that a generalized class of entropy metrics based on Sharma–Mittal entropy offers advantages over existing methods to measure design variety. We show that an exemplar metric from Sharma–Mittal entropy, namely, the Herfindahl–Hirschman index for design (HHID) has the following desirable advantages over existing metrics: (a) more accuracy: it better aligns with human ratings compared to existing and commonly used tree-based metrics for two new datasets; (b) higher sensitivity: it has higher sensitivity compared to existing methods when distinguishing between the variety of sets; (c) allows efficient optimization: it is a submodular function, which enables one to optimize design variety using a polynomial time greedy algorithm; and (d) generalizes to multiple metrics: many existing metrics can be derived by changing the parameters of this metric, which allows a researcher to fit the metric to better represent variety for new domains. This article also contributes a procedure for comparing metrics used to measure variety via constructing ground truth datasets from pairwise comparisons. Overall, our results shed light on some qualities that good design variety metrics should possess and the nontrivial challenges associated with collecting the data needed to measure those qualities. [DOI: 10.1115/1.4048743]

Keywords: design metrics, creativity and concept generation, design evaluation, design theory and methodology

Introduction

Creativity is the capacity to generate unique and original work that is useful [1–3]. Creative solutions help individuals in solving day-to-day tasks and societies by yielding meaningful scientific findings [2].

Past research [4] 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 [5–7]. Prior studies support that chances of solving a problem increase when a more diverse set of ideas is produced in the initial stages of the design process [8–10]. These findings encourage the need to explore the design space in the early stages of design [11]. But how does one quantify design space exploration?

Engineering researchers have sought to capture how "explored the solution space" is by measuring the design variety [8]. Variety, defined as the measure of how much of the solution space is covered, is a measure of ideation effectiveness since exploring the highest number of solutions in the idea generation phase is critical [8] and leads to more mature solutions to the design problem [12]. In addition, considering a variety of solutions calls for restructuring the design problem [8], allowing designers to look at a problem from different perspectives. There are two approaches typically deployed in the engineering literature to measure design variety: subjective and objective ratings of variety. As one example of subjectively evaluating design variety, Linsey et al. [13] proposed taking a set of ideas and dividing them into pools based on intuitive categories created by the coder. After this sorting process, an individual's variety score is determined by counting the number of bins into which their ideas were sorted and dividing that number by the total number of bins. The metric relies on a rater's mental model rather than a numerical procedure. 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 the reliability of the metric [14]. This article does not investigate this class of subjective variety metrics.

In contrast to subjective ratings, the other approach to measure design variety is using an objective approach (typically implemented with some type of numerical procedure using codes provided by raters using a rubric). Within those approaches, genealogical tree approaches are widely used to measure variety, as evident by hundreds of studies citing them [8,15,16]. In these approaches, subjective human raters are replaced with a deterministic formula that depends on a few measured attributes of a set of designs. One of the first metrics to use this approach was developed by Shah et al. [8] (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, researchers have reported a lack of sensitivity and accuracy of SVS [17–19]. For example, Linsey [17] showed that the genealogical tree calculation method (like SVS) is inconsistent with experts' ratings of variety. Besides, studies have shown that the sensitivity of the SVS metric diminishes when it is applied to large datasets [18] due to the exclusion of important abstract differences and generally focuses on

[1]Corresponding author.

Contributed by the Design Automation Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received February 19, 2020; final manuscript received September 3, 2020; published online November 18, 2020. Assoc. Editor: Jitesh H. Panchal.

dissimilarity in the embodiment level [19]. Alternatively, one can bin concepts via other quantitative criteria provided by structured rubrics, such as by counting the number of requirements satisfied by the designs [9]. As explained later in this article, our approach generalizes most bin counting procedures for variety measurement by appropriately setting the order parameter of Eq. (3).

While computing a variety metric score for a set of ideas is straightforward, finding which subset of ideas has the highest variety score often requires computing the score for all possible combinations. For large datasets, this process becomes computationally expensive for any variety metric, which does not have a computationally tractable method of optimization (for example, more than one billion SVS evaluations would be needed for finding 6 of 100 ideas with the highest SVS score). Finally, different metrics may be more suitable for different domains. However, there is a lack of understanding of connections between these metrics, which are measuring the same underlying phenomenon of variety in a domain. While Fuge et al. [20] argued that a few variety metrics belong to the same family of mathematical functions, they did not propose a single parametrized function that can unify many variety measurement methods.

This article argues that a generalized class of entropy metrics based on Sharma–Mittal entropy (SME) offers advantages over existing methods to measure design variety. It reexamines two existing variety metrics and compares them to methods of calculating diversity from other (nonengineering) domains. Specifically, this article compares the tree-based metrics of SVS [8] and NM [15] with the entropy-based measure of the Herfindahl–Hirschman index (HHI), which belongs to the broader class of SME metrics. This article shows how to adapt HHI to engineering design problems and proposes a new metric named Herfindahl–Hirschman index for design (HHID). By comparing HHID with SVS [8] and NM [15], this article argues and empirically demonstrates that HHID is a more accurate and sensitive measure for variety that has clear benefits for engineering and design measurement applications. We also demonstrate that this new metric is optimizable and can be generalized using a broader class of metrics.

The key contributions of this article can be categorized under five themes—accuracy, sensitivity, optimizability, generalizability, and a ground truth dataset:

(1) Accuracy: This article proposes a measurement procedure that can estimate the accuracy of variety metrics via alignment with ground truth datasets[2] comprising pairwise comparisons. Results 2 and 7 discuss how they are established using pairwise comparisons. Any metric that gives the same relative scores as a query in this dataset is considered accurate. Using a new family of variety metrics, the findings indicate that entropy-based metrics better align with human judgments of variety compared to two existing tree-based metrics for two datasets used in this study.
(2) Sensitivity: This article proposes a method of approximating metric sensitivity by randomly selecting sets and comparing their scores. The analysis shows that the SVS and NM metrics give the same variety score to a large percentage of sets (approximately 30% for our dataset), while the HHID index has higher sensitivity in distinguishing between different sets of ideas.
(3) Optimizability: The metric functions proposed in this article are monotone nondecreasing and submodular,[3] which allows one to propose a scalable greedy optimization algorithm with a constant factor optimality guarantee. To find a set of five designs with the highest variety from a collection of 1000 designs, brute force using traditional metrics makes more than 8 trillion metric evaluations, while greedy optimization gives the near-optimal solution in less than 5000 metric evaluations. This represents an efficiency improvement of around six orders of magnitude for even just a modest-sized problem.
(4) Generalizability: This article proposes that a general class of entropy-based metrics based on Sharma–Mittal entropy can be used to measure variety. We discuss how the choice of two parameters in the Sharma–Mittal entropy family affects the type of variety one wants to measure. This enables one to customize the behavior of the variety metric to a broader set of behaviors that current variety metrics can model.
(5) Ground Truth Dataset: The study leads to two datasets of pairwise comparisons that are released for future researchers to use. These datasets with pairwise queries can be used as a common scale to measure improvement in variety metrics in future studies.

The proposed family of metric has a few limitations. First, our experiments were limited to two datasets. Second, when different attributes have hierarchical relationships (e.g., an electric motor is dependent on electricity as the mode of power), the proposed metrics do not model these relationships while calculating variety.

Background and Related Work

This section first reviews some qualities that good variety metrics should possess. Then, it discusses what factors researchers should consider when constructing a ground truth evaluation method for comparing variety metrics.[4] Finally, it reviews the existing design variety metrics literature.

What Qualities Should a Good Design Metric Possess? Quality control is essential when creating and evaluating metrics that map abstract concepts like creativity to numerical scores. Particularly when metrics can be either subjective or objective, researchers needs to demonstrate that they are valid and reliable without circularity [21]. Design metrics can be relative or absolute. Relative design metrics compare ideas against other ideas in the same generated set [22]. In this way, designs generated in the same design session addressing the same problem can be compared and contrasted to tease out designs to develop further. In contrast, absolute metrics are not dependent on what other ideas are in the set. Researchers also need to reduce the subjectivity in measurement techniques, so the results do not depend on individual judges. 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 minors changes to the test environment or experimental setup [23]. However, these questions only ensure repeatability and not validity. Validity refers to the extent to which a measurement reflects the absolute state of an artifact under observation—that is, a ground truth. The term "valid" refers to an external frame of reference or a universally accepted standard against which a measurement is tested [24]. Many creativity metrics leverage a rater's expertise in a given domain to ensure metric validity [25]. This is necessary to eliminate circularity or measuring unvalidated metrics against other un-validated metrics [26]. In this article, the term "accuracy" is used to measure the validity of a metric against an established standard, which we call the ground truth.

The key assumption in many past works is that raters who have considerable experience in a given domain are best suited to provide the ground truth assessment for tasks like evaluating creativity [27]. If experts are the de-facto ground truth, then why do we need a separate, objective metric? This is because of resource and practical constraints: expert time and effort is a scarce commodity. This scarcity forces researchers to develop objective metrics that can aid quasi-experts or novice raters in accurately evaluating processes

[2]We use the term "ground truth dataset" to refer to a set of design examples, where one is confident of the variety measurement. These measurements can be derived from expert feedback, domain knowledge, or consensus from many people.
[3]Submodular functions are set functions to model diminishing marginal utility.
[4]We use the term "variety metrics" for metrics used to measure design variety.

and ideas. But how do we verify whether a proposed metric is valid? This article focuses on how to validate any proposed objective metric against expert raters. This article focuses on how variety metrics must be evaluated to ensure they are measuring what they are built to measure, reliably, and with an acceptable degree of validity.

When a metric is created, it is important to establish some desiderata (qualities we want) that a metric must possess. The prior work on establishing acceptable qualities of a metric includes the work by Amabile and Pillemer [28], who were key in standardizing the measurement of creativity in psychological research. Previously, most methods used pencil and paper tests, personality tests, biographical inventories (such as Schaefer and Anastasi's biographical inventory [29] and Taylor's Alpha Biographical Inventory [30]), etc. These tests were debatable in experiments that sought to reduce within-group variability and generally lacked a clear creativity definition and an effective strategy to avoid biases on behalf of the rater [28]. Amabile and Pillemer's work highlighted the need to better understand the multiple desiderata for a creativity metric. Building upon that work, this article attempts to mathematically describe and lay out experimental procedures by which one might measure such desiderata.

For example, good metrics should have the ability to establish ground truths using expert agreements and must be replicable by other raters who use the metric. In this regard, variety metrics like SVS and NM were developed to reduce subjectivity on the rater's part and make it easier for researchers to replicate processes used to analyze designs. For subjective metrics, Cropley [31] argues that high interrater reliability and internal consistency are desirable metric qualities.

This article argues that for any new design variety metric, accuracy, sensitivity, repeatability, explainability, and optimizability are also desirable qualities. Here, accuracy means that if ground truth estimates of a quantity are available, then a new metric should align with this ground truth. Sensitivity means that a new metric should be able to distinguish changes between different states of a quantity. Repeatability means that when measurements are repeated again and again with the quantity being unchanged, they should not give different measurements. Explainability means that the measuring instrument should give explainable scores, that is, it should be possible to explain why one set of designs received a higher score than another set. Finally, optimizability means that given a ground set of ideas, if the goal is to find sets of ideas that will have maximum or minimum measurement score using a variety metric, then practitioners should be able to do so in polynomial time (where time is a simple polynomial function of the size of the input, for example, the number of designs considered). Subjective metrics generally lack repeatability and explainability. In contrast, existing metrics like SVS and NM are repeatable and explainable. However, this work shows that SVS and NM are not accurate, sensitive, and optimizable compared to the Sharma–Mittal family of metrics.

Why is Measuring Design Variety Important? Engineering researchers introduced design variety metrics to measure how well someone explores the solution space during a design task [32]. Generating a large number of ideas with iterative or small changes may not result in effective concept generation or innovative products. Research has shown that "there is no way to generate an optimum solution without exploring the solution space through early tentative ideas" ([33]), which shows the importance of measuring design variety. 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 [8]. Cognitive restructuring is frequently used in concert with the number of ideas developed (quantity) to assess design ideation.

Research in engineering design has shown a correlation between the amount of design space explored and the quality of the final design [34]. In engineering design education too, variety, and number of ideas generated in the concept generation stage was highly correlated with students' performance on the design project [35]. Researchers have found that the consideration of a variety of solutions is important to provide a cognitive restructuring of the design problem [36]. As a result, the variety of the solution space is highly correlated with the novelty of the idea set [37] and the quality of the final product [34] leading to more mature solutions to the design problem [12]. Without exploration, designers may misconstrue the solution space to be narrow. One of the main contributing factors to this trend is functional fixation, or blind adherence to solutions that are familiar and comfortable, which can generally lead to products of lower quality or innovation [38,39].

Providing examples have been found to help designers explore the solution space and drive design innovation [40] by providing jumping-off points for designers [41]. Suppose a design practitioner conducts an ideation exercise with ten teams, each of which generates five designs. The design practitioner wants to select five designs from the 50 generated designs to use as an inspiration for future participants of ideation exercises. Ideally, the selected designs should be high-quality and cause minimum functional fixation. How should one do this? To choose such a set objectively and know how much a design space is explored, the design practitioner needs to measure both the quality of the designs and their variety. In this work, we focus only on the measurement of design variety as quality measurement is often domain dependent.

Measuring design space exploration requires computing mathematical functions on groups of ideas [14]. To address this need to measure the extent to which tools promote variety, Shah et al. [8] developed a metric (SVS) to provide a repeatable and reliable method to calculate design variety by rewarding ideas that are differentiated at higher levels of abstraction. In the SVS metric, the authors decompose design variety into four hierarchical levels: the physical principle, followed by the working principle, embodiment, and detail.

Researchers have found that the SVS metric double counts ideas at each level in the tree, and there is a lack of guidance on how the specific numerical choice of the weights at each level of the tree is to be determined [15,42]. Because of these pitfalls, Nelson et al. [15] 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. Besides, Nelson et al. [15] modified the SVS metric by altering the weighting scheme from 10, 6, 3, and 1 to 10, 5, 2, and 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.

However, both SVS and NM do not define what each level of the hierarchy means. There has been insufficient empirical justification or verification of the weights used in such genealogical tree metrics [17], which can lead to large variations in the deployment of the metric in engineering design research. Srinivasan and Chakrabarti [43] also propose an idea space variety metric and base it on the abstraction levels of the SAPPhIRE model, with different weights for action, state change, input, phenomenon, effect, organ, and part abstraction levels. Other improvements of the SVS metric include the work of Verhaegen et al. [16], who combined Shah's metric with a Herfindahl index-based tree entropy penalty, to encourage "uniformness of distribution"—essentially preferring trees that have even branching. Verhaegen et al. [44] subsequently showed problems with many tree-based metrics, including arbitrarily defined weights. However, most of these methods require constructing a hierarchical tree. Our analysis demonstrates that the additional step of constructing a tree may not be necessary for measuring design variety for many domains as entropy-based metrics, which do not require tree construction often align better with human assessment of variety.

Apart from using hierarchical trees, researchers have employed many other variety metrics of varying complexity. For instance,

one of the measures of variety defined is the ratio of the number of categories that a participants' ideas occupied to the total number of bins [13,45]. Later, we show that the generalized two-parameter Sharma–Mittal entropy metrics reduce to a form very similar to the above metric by selecting both parameters to zero. Thevenot and Simpson [46] proposed a comprehensive metric for commonality to evaluate the design of a product family based on product attributes and the allowed diversity in the family. Henderson et al. [11] compared different variety metrics and proposed a new metric that calculates variety by looking at how a collection of ideas covers a potential design space based on the diversity of the other metrics used to assess those ideas. While most metrics discussed so far focused on measuring variety in an ideation exercise, Kota et al. [47] proposed product line commonality index, to capture the level of component commonality in a product family, which was later used in an integrated platform by Jung and Simpson [48] to support product family redesign. While many of the variety metrics discussed earlier have shown promising results in different domains, there is a lack of methods to combine metrics sharing common components or to find a unifying formulation, which connects variety metrics with theoretical concepts like the entropy of a system. There is also a lack of criteria that a new variety metric should satisfy. This article tries to address some of these gaps.

Measuring Variety in Other Domains. Variety metrics are used in different domains like economics and ecology under different names. They are often referred to as diversity metrics, while terms like coverage, breadth, or heterogeneity are also used. For ideation, researchers have measured the breadth of ideation using metrics like the mean pairwise distance between ideas [49] or by manually subgrouping functions into categories [50]. Over the last 20 years, economists have also become increasingly interested in understanding whether diversity among multiple distinct population groups enhances or impedes a society's economic and social development. To quantify the economic impact of diversity, they also needed to create an index that captures how society divides into various factions or parts.

Starting from the Gini index [51], economists have used various diversity indices to evaluate the degree of social, economic, cultural, and other dissimilarities among people, regions, and countries. The Gini index was re-interpreted by Simpson [52] as the inverse Hirschman–Herfindahl index. A variety of other statistical metrics of diversity including Shannon entropy [53], effective numbers of species (aka Hill's metric), Tsallis number, etc. are also commonly used in many fields including information theory (to measure the amount of information conveyed) and ecology (to measure diversity of species). The following paragraphs discuss three of these measurement methods—Shannon entropy [53], Richness [54], and HHI [55].

The most commonly used diversity metric is called Shannon Entropy. Shannon entropy quantifies the uncertainty in predicting the group identity of an individual item that is taken at random from the dataset. Shannon entropy becomes zero when there is exactly one group, that is, there is no uncertainty in predicting the type of the next randomly chosen item.

Richness quantifies how many different types of categories the dataset of interest contains and is a popular diversity index in ecology. Although widely used, richness does not take into account the abundances of each type within their group. On careful observation, one may notice that this same problem occurs if one is using design metrics inspired by the "Richness" metric, which count the number of bins in a set of designs [13]. This property (called evenness sensitivity in Ref. [56]) is satisfied by other metrics like Shannon entropy, which consider the abundances in each category and the number of categories. It is also interesting to note connections of richness variety metric with Shah's novelty metric. If a set of designs has all novel ideas, one would expect that each design in this set does not share its bin with any other design. Hence, each item will be unique (as measured by Shah's definition of novelty), and it will also be diverse (as measured by the richness metric).

The HHI is a statistical measure of concentration [55,57]. HHI is used by the Department of Justice and the Federal Reserve in the analysis of the competitive effects of mergers. It accounts for the number of firms in a market, as well as their concentration, by incorporating the relative size (i.e., market share) of all firms in a market. For a market with N firms, HHI is calculated by squaring the market share (MS_i) of all firms ($i \in \{1, \ldots, N\}$) in a market and then summing the squares as follows:

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

Markets with more concentration (less variety) will have a few large square terms. HHI has also been used in other domains ranging from the measurement of linguistic diversity [58] to the measurement of academic specialization [59]. This article proposes a variant of the HHI metric named HHID. The metric does not necessitate finding hierarchical trees, simplifying the variety calculation. The results show that unlike past metrics, this new metric also has better alignment with the judgment of variety by people.

While studying HHID, one may ask why should someone use HHID and not another metric, which is a slight variant of it (say cubic power instead of square terms)? To answer this, we show below that the HHID metric is just one instance of a more generalized class of Sharma–Mittal entropy metrics, which can be used to measure design variety. Commonly used Hartley, Shannon, and Quadratic entropy, and the families of Tsallis, Renyi, and Arimoto entropies, can all be derived as special cases of Sharma–Mittal entropy metric [60]. This insight helps unify past notions of design variety under a common mathematical form.

Unifying the Space of Variety Metrics. SME is a generalized class of entropy measurement methods that unifies multiple past proposals to measure diversity. It argues that the uncertainty in a discrete random variable $K = k_1, k_2, \ldots k_n$ can be measured by its entropy. SME can be defined as follows:

$$\text{SME}(K) = \frac{1}{t-1}\left[1 - \left(\sum_{i=1}^{n} P(k_i)^r\right)^{(t-1)/(r-1)}\right] \tag{2}$$

where r is the order and t is the degree of the entropy measure. The order r is any positive real-valued number except 1. The degree t can be any real-valued number except 1. $P(k_i)$ is the proportion of ideas of variable k. Figure 1 shows how metrics like Shannon, Quadratic (HHI), Tsallis, Effective number, and others can be obtained using different values of orders and degree parameters.[5]

Although the Eq. (2) may not immediately appear intuitive, there are many ways to build an understanding of this space of metrics. For example, all of the SME metrics can be thought of as quantifying the average surprise that would be experienced if the value of the random variable K was learned. The order parameter r determines what kind of averaging function is used. r can be thought of as an index of the imbalance of the entropy function, which indicates how much the entropy measure discounts minor (low probability) hypotheses. For example, when $r = 0$, entropy becomes an increasing function of the mere number of the available options. When r goes to infinity, on the other hand, entropy becomes a (decreasing) function of the probability of a single most likely hypothesis. The degree parameter t governs which kind of surprise is averaged. It can be considered as a deformation parameter of the probability distribution [61], and unlike r, it does not have an intuitive explanation. While these relationships between r and t may, at first, appear to just be mathematical curiosities, we show below that by viewing variety in this way, researchers can better scientifically study and uncover how people

[5]Note that limits, which exist, are used for points where Eq. (2) is undefined.

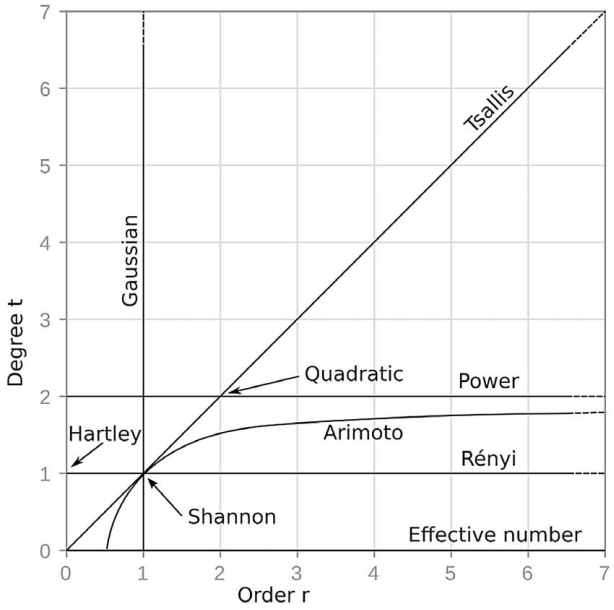

Fig. 1 The two-parameter Sharma–Mittal entropies. Different existing entropy metrics like Shannon, Hartley, Tsallis, Quadratic, etc. are incorporated in this class of entropies.

make decisions about variety—for example, by determining ranges of r and t that agree well with expert opinion.

Methodology

In this section, we first describe variety measurement methods using the Sharma–Mittal entropy and then show how a Herfindahl–Hirschman index-based metric can be derived from it. Next, we show an example of variety calculation 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. We finally show example computation of variety using the Sharma–Mittal entropy, which generalizes HHI.

The Sharma–Mittal Entropy for Design. In this section, we propose a variant of SME that can measure the variety of a set of designs. To do so, we assume that we are given a set of designs S. As commonly used in the literature, it is assumed that a design is represented by a certain level of abstraction like the physical principle, the working principle, the embodiment, and the detail level. As explained by Verhaegen et al. [44], a generated concept of a motor could, for instance, exist of the ideas "electromagnetism" at the physical principle level, "coils for attracting and repelling permanent magnets" at a working principle level, a schematic or description of the placement of the coils and permanent magnets on the shaft and casing at the embodiment level, and a detailed drawing or description of the parts and assembly at the detail level.

Each design within a set S can be described by a list of attributes (the attributes can be hierarchical levels like functional principle, working principle, embodiment, and detail similar to SVS and NM above or they can be nonhierarchical categorical attributes). We define Sharma–Mittal entropy for design (SMED) for each attribute by replacing $P(k_i)$ in Eq. (2) by the corresponding proportion of functional principle. Hence, $\text{SMED}_F(S)$ for functional principle is defined as follows:

$$\text{SMED}_F(S) = \frac{1}{t-1}\left[1 - \left(\sum_{i=1}^{N_f}\left(\frac{|FP_i|}{N}\right)^r\right)^{(t-1)/(r-1)}\right] \tag{3}$$

where $|FP_i|$ is the number of designs using functional principle i and N_f is the total number of functional principles (or the number of categories based on any factor, as defined by a designer). N is the total number of designs in the set S. Similarly, the article defines SMED_W for working principle, SMED_E for embodiment, and $SMED_D$ for details (or any number of attributes defined for a design). The total variety score for a set S is defined as the weighted sum of variety score for each type of attribute as follows:

$$\begin{aligned}\text{SMED}(S) = w_1\ \text{SMED}_F(S) + w_2\text{SMED}_W(S)\\ + w_3\text{SMED}_E(S) + w_4\text{SMED}_D(S)\end{aligned} \tag{4}$$

Here, $\text{SMED}(S)$ is the total variety score for a set of designs S. The weights w_1, w_2, w_3, and w_4 are used to give different importance to the variety of different attribute types and can be set such that the resultant value is always bounded to be less than one (say, by setting the sum of weights to 1). For instance, if all factors are equally important, then one can set $w_1 = w_2 = w_3 = w_4 = 1/4$. This article assumes here that the total variety of a set is a weighted linear sum of the variety of different attributes found in that set. As discussed later, this assumption aligns well with human judgments and also allows the resultant metric to remain submodular under certain conditions. By varying the parameters of SMED, one can measure different types of variety. For instance, if one selects $r = t = 0$, the metric reduces to $\text{SMED}_F(S) =$ number of unique attributes-1, where the attributes or categories can be the functional principles or subjectively defined categories by an expert. This reduction gives a metric, which is similar to the metric proposed by Linsey et al. [13], which counts the proportion of unique bins (categories) to the total number of bins. The next section shows how HHI defined in Eq. (1) is a special case of SMED metric defined in Eq. (3) by using $r = t = 2$. These specific values of r and t are selected as they are the smallest integral values satisfying the optimizability criteria. HHI is a common measure used in domains like economics, and as shown by Ahmed et al. [62], it aligns well with human interpretation of variety.

The Herfindahl–Hirschman Index for Design. The Herfindahl index (also known as Herfindahl–Hirschman index) measures a firm's size relative to the industry and indicates the amount of competition among firms. The mathematical structure of HHI was provided in Eq. (1). The value of HHI measures the probability that two randomly chosen individuals in society belong to the same groups.[6] This section proposes a variant of HHI, named HHID, that can measure the variety of a set of designs, where each design is described by a set of attributes. We calculate the HHID index for each attribute separately for the entire set. For example, the HHID index for the "functional principle" attribute type is given by

$$\text{HHID}_F(S) = 1 - \frac{\sum_{i=1}^{N_f}|FP_i|^2}{N^2} \tag{5}$$

One may notice that $\text{HHID}_F(S)$ can be derived from $\text{SMED}_F(S)$ in Eq. (3) by setting both r and t parameters to two, showing that HHID is a special case of the broader Sharma–Mittal class of metrics. When $N_f \geq N$, $\text{HHID}_F(S)$ varies between 0 and $1 - 1/N$. Unlike the HHI definition in Eqs. (1) and (5) subtracts the value from 1. This definition is closer to the Gini–Simpson index, which is also known in ecology as the probability of interspecific encounter [63]. $\text{HHID}_F(S)$'s value is maximum when all ideas have unique functional principles in the set. Mathematically, it measures the probability that two randomly chosen ideas in the set have different functional principles. Similar to SMED, the total variety score for a set S can be defined as the weighted sum of variety score for each type of an attribute as follows:

$$\begin{aligned}\text{HHID}(S) = w_1\text{HHID}_F(S) + w_2\ \text{HHID}_W(S)\\ + w_3\text{HHID}_E(S) + w_4\text{HHID}_D(S)\end{aligned} \tag{6}$$

[6]The interpretation of HHI as the probability that two individuals selected at random from a set represent the same group assumes that the first person is replaced with the set before taking the second person.

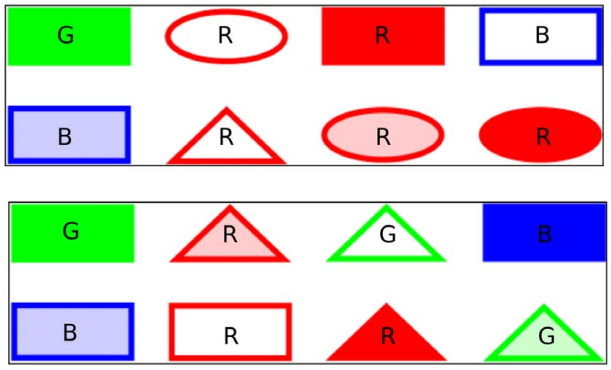

Fig. 2 Example of two polygon sets (top shows set A and bottom shows set B) shown to participants in our experiment. Participant answers the question: "Which set has higher variety?". Note that the letters R, B, and G are added to indicate red, blue, and green colored shapes for improved readability of this article. These letters were not shown to the participants.

Example Variety Calculation Using Proposed Metrics. To demonstrate HHID calculation, an illustrative example is discussed next, which is shown in Fig. 2 with two sets of items. In this example, we use polygons instead of a case study from engineering design due to two reasons—attributes like shape and color are easy to visualize and one can do one to one mapping of a polygon attributes to the attributes of any engineering design idea with a similar number of total attributes.

In Fig. 2, 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 shape. There are five red-colored polygons (marked by R), two blue (marked by B), and one green (marked by G). Three items have a solid fill, two have shaded, and three are empty inside. Without the loss of generality, for example, we assume that color is the functional principle of a polygon, shape is the working principle, and shading is the embodiment. It is also assumed that all three levels are equally important in deciding the variety of set A ($w_1 = w_2 = w_3 = \frac{1}{3}$) and $N_f = 3$ as there are three unique functional principles (color). The HHID_F score for color will be $1 - ((5/8)^2 + (2/8)^2 + (1/8)^2) = 0.531$. Similarly, HHID_W score for shape will be $1 - ((4/8)^2 + (3/8)^2 + (1/8)^2) = 0.593$ and HHID_E score for fill will be $1 - ((3/8)^2 + (2/8)^2 + (3/8)^2) = 0.656$. As all features are assumed to be equally important, the total HHID for the set of designs will be $(0.531 + 0.593 + 0.656)/3 = 0.593$. Similarly, the variety of any set of designs can be calculated. For instance, the HHID(S) score for the set at bottom is $(0.656 + 0.500 + 0.656)/3 = 0.604$, using Eq. (6). These two sets are close in their variety scores using HHID, with the bottom set having a slightly higher score than the top set. While the top set lacks in color variety (five red polygons), the bottom set lacks in shape variety (no oval shape). The above scores also demonstrate that we can introspect on the metric or designs to uncover or explain why some sets score higher than others. This is in contrast to subjective expert-provided "black box" type methods, which do not have an easy method to invoke queries related to explainability. By using HHID, one can calculate sensitivities of the metric to new features. In future work, we will explore what groups of feature vectors are responsible for a higher variety of a set.

Next, the variety score of both the sets using SMED is calculated. For demonstration, we use two settings of order and degree parameters. First, both values are set to zero, i.e., $r = t = 0$. In this case, the SMED_F, SMED_W, and SMED_E are each one minus the number of unique principles (three). Hence, the SMED(S) is six for the top set. The bottom sets SMED(S) is five (as it lacks one type of shape). This shows that the top set has a higher variety score if SMED parameters are set such that they give more importance only to the unique number of groups found. The second setting we show is $r = 4$ and $t = 2$. The SMED(S) scores for the top and the bottom set are 0.558 and 0.598, respectively, again giving a higher variety score to the bottom set, similar to HHID. In all these cases, it is assumed that variety for color, shape, and shading are equally important. However, suppose a problem requires that the shape variety is four times more important than color and fill, then we can set $w_1 = w_3 = \frac{1}{6}$, and $w_2 = \frac{4}{6}$. In that case, we get variety score of 0.561 and 0.549 for the top and the bottom set, respectively. The first set, which has three unique shapes, gets a higher variety score if more importance is given to the variety of shape. Thus, by changing the weights for different attributes, one can customize the variety metric to meet the demands of a particular domain.

Optimizing Variety of a Set. Using metrics like SVS, NM, and HHID, one can measure the variety of a given set of ideas (like the sets shown in Fig. 2). However, what happens when one wants to choose a small set of polygons (say five) that have the maximum variety out of a thousand items? One way is to enumerate all possible sets of size five (more than 8 trillion sets for a ground set of 1000 items), calculate their variety score, and then find the set that have the highest variety score. This approach becomes intractable as the number of items in the ground set increases.

Another approach, and the one used in this article, is to leverage mathematical properties of the variety function and find approximate solutions close to the optimal. This HHID metric is a submodular set function. Submodular functions are functions defined over sets that are designed to model diminishing marginal utility, which is the mathematical property needs to model diversity or variety [20]. Having the submodularity property means that the variety metric follows the law of diminishing returns—when a design is added to a larger set, the increase in HHID score is smaller compared to the case when the same design is added to a smaller set. This property can be exploited to find sets of maximum variety using a greedy algorithm [64], which guarantees that the variety of the greedy search solution will be within 63.2% (or $1 - (1/e)$) of the variety of the optimal solution.

To find sets of maximum variety, one can use a submodular greedy algorithm explained in Ref. [64]. Given the set V of all ideas, the algorithm starts with an empty set $S = \{\}$ and add ideas to this set, which give the maximum marginal gain in the submodular function. At every step, it adds one idea at a time, such that the selected idea $i \in V$ is the one with the highest marginal gain $\delta\text{HHID}(S \cup i)$ on set S. At each step, the algorithm adds the idea that will give a maximum increase in variety in the set S. Finally, as the function in Eq. (6) is submodular and monotonic, the algorithm is also theoretically guaranteed to provide the best possible $(1 - (1/e))$ polynomial time approximation to the optimal solution [65,66]. $\text{SMED}_F(S)$ function defined in Eq. (4) is concave as long as $t \geq 2 - 1/r$ (Hoffmann et al. [67] provide a proof). As a sum of concave functions over modular functions is submodular [68], the resultant SMED(S) metric is also submodular for $t \geq 2 - 1/r$. Hence, the key takeaway is that one can optimize any SME derived metrics (like HHID) for all values $t \geq 2 - 1/r$ in polynomial time using a simple greedy algorithm. We later use this property to show saving in computational time to find highest variety sets.

Experiments and Results

We conducted an experiment to benchmark the proposed HHID metric with the commonly used SVS and NM metrics using a known and easily verifiable ground truth based on polygons. Next, another experiment is reported, which uses milk frother design sketches provided by engineering students and rated by domain experts. Before introducing our experiment and its main results and implications, we describe how the experimental dataset of set comparisons was constructed. As shown later, constructing such sets is nontrivial, and one contribution of this

article 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 30 design ideas has more than one billion possible sets (2^{30}) of designs for which variety needs to be calculated. Exhaustively calculating the ground truth for all designs is infeasible. To avoid circularity, any existing variety metrics are not used to create the ground truth. Doing so would assume that a given metric represents the true variety, which is what the ground truth is used to establish. Instead, this article proposes the development of a ground truth by directly asking human raters. To establish a ground truth dataset for calculating the design variety, three components are needed:

(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 to enable the calculation of tree-based metrics

Variety scores are calculated on a set of designs. However, human raters are not good at giving absolute scores [69] due to differences between internal scales of subjects. For instance, given the set of designs shown in Fig. 5 (top), it would be difficult for a human rater to say whether this set of six designs scores 6 of 10 or 8 of 10 for variety. Different raters may also use different internal scales.

In contrast, if a rater is asked to rate whether they find the variety of set shown in Fig. 5 (top) greater than the variety of those shown in Fig. 5 (bottom), they may answer it relatively easily because humans are better at comparing items than giving absolute scores [70]. Hence, this article proposes that a ground truth dataset for variety should be created using pairwise queries (ordinal judgments), where each query contains two sets and there is a consensus among human raters that one set has higher variety compared to the other set. To elicit responses from experts, two sets at a time are given to them and they are asked pairwise comparisons of the form: "Which set of designs has higher variety?"

Measuring Variety for Polygons. In this experiment, the performance of SVS and NM metrics in measuring the variety of a set of polygons is compared with HHID, which is a special case of Sharma–Mittal entropy. Initially, a base set of 27 polygons is created. 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. These colors are indicated in the figures by alphabets R, B, and G, respectively. Shading varies between polygons as complete fill, shaded, or empty.

The polygon example, which does not represent a real-world design, is intentionally chosen to compare design metrics. The difficulty with using a real-world example to establish a ground truth for objective metrics is that such examples have many moving parts and human judges have low agreement on what attributes should be extracted from the design and which ones are important in determining their similarity. For real-world examples, due to the inherent complexity in the measurement of attributes and design performance, it is difficult to say conclusively say whether the lack of alignment of human judgment with a variety metric is due to the wrong choice of attributes or the wrong choice of the method measuring variety over those attributes. While the polygon example does not represent an actual engineering design solution, it is used to compare metrics when there is no ambiguity in design attributes (shape, color, and fill). Our argument is that metrics that can measure variety for many complex domains should at least fair well in measuring variety for a simpler polygon-based ground truth dataset. Later sections provide a more complex example and discuss the issue with capturing attributes.

The total number of possible sets of polygons is large (2^{27}); hence, calculating the variety score of all possible sets is time consuming. Instead, the search is narrowed down to focus on three set sizes: when the number of items in a set is four, six, and eight. The researchers observed in their preliminary experiments that if human raters are asked to compare sets with larger than eight items, the task becomes too difficult for them, as evident by low agreement between different raters. For a given set size (say size six), the total number of ways two sets can be compared is also quite large (more than 43 billion set comparisons). Hence, we first randomly select 100 sets for comparison. From these 100 sets, we calculate all possible pairwise comparisons (4950 comparisons with each comparison containing two sets of size six). Next, we calculate SVS, NM, and HHID scores for all the sets in each comparison. For SVS and NM computation, the analysis assumes 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 presents the percentage of comparisons where each metric finds both the sets of equal variety. Note that SVS and NM metrics do not distinguish between a large percentage of comparisons (31.7% and 21.4% for sets of size six), while HHID gives identical scores to a much smaller percentage of pairwise comparisons (14.7% for sets of size six). This implies that existing metrics are not sensitive or discriminative to differences between sets.

Result 2: Existing Metrics Vote Similarly to One Another. Table 2 presents the percentage agreement between different metrics. 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 HHID and other metrics is close to random. Due to the lack of a benchmark dataset, it is difficult to comment on whether a lack of agreement between metrics is a good thing. We show later in the results that HHID aligns with the human raters more than SVS and NM.

Table 1 Percentage of pairwise comparisons when design metrics give same score to both designs

Method	Same score SVS (%)	NM (%)	HHID (%)
Size 4	27.3	37.0	15.8
Size 6	31.7	21.4	14.7
Size 8	28.5	12.9	10.9
Size 10	31.2	14.5	9.2

Note: Lower percentages are good as it indicates that a metric can distinguish between sets. SVS metric gives same score for approximately 30% of the sets.

Table 2 Agreement between metrics for pairwise comparisons

Method	Agreement SVS-NM (%)	HHI-SVS (%)	HHI-NM (%)
Size 4	84.4	54.2	50.2
Size 6	81.0	47.6	50.0
Size 8	82.5	49.4	56.9
Size 10	84.4	54.2	50.2

Note: SVS and NM tend to vote similarly for more than 80% of the sets.

Method to Establish a Ground Truth Dataset for Variety Metrics Assessment. Establishing the ground truth for comparing different metrics required the following steps. First, pairwise comparisons where SVS and NM could distinguish between the two sets are selected; that is, both the metrics did not calculate the same variety score for both sets. This is important since we 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, the sets where both metrics disagreed on their vote are selected. This means if SVS scored set A to have higher variety, then NM would give set B a higher variety score. 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 and tend to give same scores to a large percentage (up to 37%) of the sets.

Finally, the top 5 sets where SVS is most confident that one set has higher variety than another are selected 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 ten queries, which are then given to human raters. Finding human annotations for such sets allows a researcher to find out which of the two metrics better aligns with human responses.

To find the ground truth for polygons, an Amazon Turk study was conducted to collect responses from crowd workers for pairwise queries. A sample query with two sets of eight polygons is shown in Fig. 2. Judging the variety of polygons does not require expertise in the area, and Amazon Turk enables getting a large number of responses quickly. We collected pairwise responses for three different set sizes. For each set size, ten pairwise queries were created. For each query, ten responses from Amazon Turk participants were collected. We used total of 30 Amazon Turk workers, who took an average time of 4 min to complete a survey and were paid $1.2. A maximum time of 15 min was allocated for the task. We used the following three criteria to accept workers for the task: (1) hit approval rate greater than 90%, (2) number of hits approved greater than 100, and (3) location is the United States. On accepting the task, a worker was directed to a Qualtrics survey page, which started with the following task description:

Context

Each question has a pair of images, namely, set A and set B. Each set has few geometrical items. These items vary by three main features:

(1) Shape: They can be triangular, rectangular, or oval shaped.
(2) Color: They can be blue, red, or green color.
(3) Fill: They can be fully colored, shaded, or white inside

Requirement

For each question, you have to judge whether set A has higher variety or set B. The set with higher variety should have items, which are more different from each other, i.e., the set should be more diverse.

To ensure the quality of responses from the crowd workers, the following steps were taken: (1) The order of the queries was randomized and also the order of the options shown to different participants to reduce the possibility of any ordering bias. (2) The surveys were divided into two parts to reduce fatigue. (3) No worker was repeated across surveys, and (4) 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 four was as follows: [9, 8, 9, 7, 6, 9, 8, 6, 8, 7]. This means that for the first query, nine people of ten voted for the same set. For the second query, eight people voted for the same set as being of higher variety and overall. Similarly, for sets of size six, [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. Hence, the average agreement between raters for sets of size four, six and eight were 79%, 72%, and 71%, respectively.

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

Based on how people voted in this experiment, the maximum number of points that any metric can receive is 220—that is, if it always votes with the majority opinion of human raters. Now, suppose a metric receives 200 points in total, then we say that it has 90.9% alignment ($100 \times 200/220 = 90.9$) with human ratings.

Result 4: HHID Outperforms SVS and NM With Respect to Human Agreement on Polygon Variety. Table 3 presents the comparison between SVS, NM, and HHID for alignment with human ratings. SVS and HHID have a similar best-case performance for this dataset. By varying the weights of each functional level between 1 and 10 in steps of 1 gives a 1000 possible performance scores corresponding to each weight combination [w1, w2, w3] (this is in contrast to using a fixed combination of weight, like [10, 6, 3, 1] for SVS). The results show that HHID performs better than SVS in the median case, where the median is calculated over all the thousand weight combinations.

Table 3 presents that HHID 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, HHID aligns with 74.5% of human ratings. We find that the highest performance is obtained for many combinations of weights. At first glance, SVS also seems to perform well for the median case. However, this does not mean SVS is suitable to measure variety as we only select the queries where SVS is able to differentiate between the two sides. It is also important to recall that the comparisons were generated, 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 alignment with human responses—humans generally showed a clear preference between the variety of two sets, but SVS would be indifferent. Hence, due to better accuracy and higher sensitivity, the HHID metric outperforms both SVS and NM in alignment with human's judgment of variety.

Table 3 Comparison of design variety metrics in alignment with human ratings

Method	Median case (%)	Best case (%)	Worst case (%)	Sample optimal weights
HHID	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

Result 5: Sharma–Mittal Entropy Parameters Show That Most Entropy Metrics Align Well With Human Judgments. This section shows how SME aligns with the human perception of variety. To do so, the weights of the metric for color, shape, and shading are set to be equal, and only the order r and degree parameters t of the SME metric are varied.

r and t are varied between 0 and 10 at steps of 1 (note that r and t are not necessarily restricted to integral values). The results are shown in Fig. 3. For each combination, the resultant alignment with humans is calculated. As shown by the white region, the metric achieves the highest performance for multiple combinations of r and t, including $r = t = 2$, which correspond to HHID. This leads to the question: What does this indicate about how people think about measuring the variety of polygons?

To dive into this question, it is important to first understand the parameters of the SME metric. In the SME metric, the order parameter r is an index of the insensitivity to less abundant principles. As r increases, variety gets closer and closer to a simple (decreasing) function of one single element in the distribution, which is the relative abundance of the most common principle (most common color, shape, or shading in the set). When $r = 0$, on the contrary, variety becomes an increasing function of the plain number of principles with nonnull relative abundance (e.g., the count of the number of unique shapes, colors, or shading). This shows that the order parameter r indicates how much a variety measure disregards relatively rare principles. The role of the degree parameter t is more technical: It affects a few important metric properties, which is elaborated in detail in the literature on mathematical analysis of SME [61].

The results show that the metric top performance is indifferent to variations in r. This means some people may have focused on just the count of classes, while others may have focused on the largest class. Performance is sensitive to values in t, with a decrease in performance as t goes above two.

Result 6: We Can Find Sets of Designs With Highest Variety. One of the auxiliary outcomes of using an HHID-derived index for variety measurement is that it provides a simple method to find the highest variety sets. Suppose in an ideation exercise, ten teams get together to generate a total of 27 ideas. Our goal is to combine all ideas into one large set and then down-select to a small set of ideas that provide a distribution over the design space. To demonstrate the concept, we assume that ideas produced by all the teams are represented by the 27 polygons discussed before, and the goal is to find a subset of five ideas, which have the highest variety (one can pick any size of the subset).

If one wants to find the subset of size five ideas with the highest SVS variety (or any other tree-based variety metric), they will have to calculate all possible combinations of five ideas, then calculate the tree for each subset, estimate SVS scores for each set, and finally pick the set with the highest SVS score. This exercise will require enumerating all 80, 730 (27 choose 5) possible trees for each set of five polygons. This approach becomes infeasible when the ground set becomes large due to a large number of possible options (mathematically, this is because the problem is NP-Hard).

In contrast, we use a greedy algorithm [64,71] to rank order all polygons or to select a subset. For size 5, it will require only 125 evaluations, which requires 99.84% fewer calculations. When applied to polygons, the resultant set, with the highest variety for color, shape, and shading, is shown in Fig. 4. The method selects one polygon at a time based on which polygon provides the highest marginal gain. A practitioner may also wonder how many ideas should provide sufficient coverage over the design space. While this work does not show how many ideas are enough to explore the design space, past work by Ahmed et al. [71,72] have shown straightforward methods to estimate the cutoff for the number of ideas needed using the marginal gain of a submodular function. The same method applies to the current variety metric due to its submodularity and can be used to find the size of the subset.

Measuring Variety for Milk Frother Sketches. While the polygon example discussed in the previous section helped to validate the metrics, using a problem with little complexity, it did not represent an actual engineering design solution. In this section, an additional experiment of an engineering design problem is discussed, where the goal is to measure the variety of early-concept design sketches. We use experts to judge items from a preexisting dataset of milk frother sketches to create a ground truth dataset comprising pairwise comparisons. Finally, we measure how well different variety metrics align with the ground truth to measure their accuracy.

To measure the variety of milk frothers, data from a previous experiment conducted by Starkey et al. [73] are taken, which consisted of 934 idea sketches. Specifically, the dataset 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 calculate the metrics based on hierarchical features, the results from the previously developed design rating survey (DRS) was used to classify the features addressed by each design concept (see the study by Toh and Miller [74] for more details). Twenty questions on the DRS were used to help raters classify the features each design concept addressed. The results of the DRS were then split into which category they addressed in the extension metrics: physical principle, working principle, or embodiment. The physical principle was determined by what type of power source was used to power the product (i.e., manual, battery). Conversely, the working principle was determined by what type of motion was used by the product (i.e., stirring, shaking) and the embodiment was determined by what the product looked like (i.e., shake weight, handheld frother).

In this study, to create the dataset of sets of milk frother sketches, a ground set of ten design sketches is adopted from the study by Ahmed et al. [75]. The benefit of using these ten sketches was the availability of hierarchical features as well as information in the form of subjective idea maps, which is later used for gaining additional insights. The total number of possible sets, which can be formed using these ten sketches is $1024(2^{10}$ *sets*). Similar to the

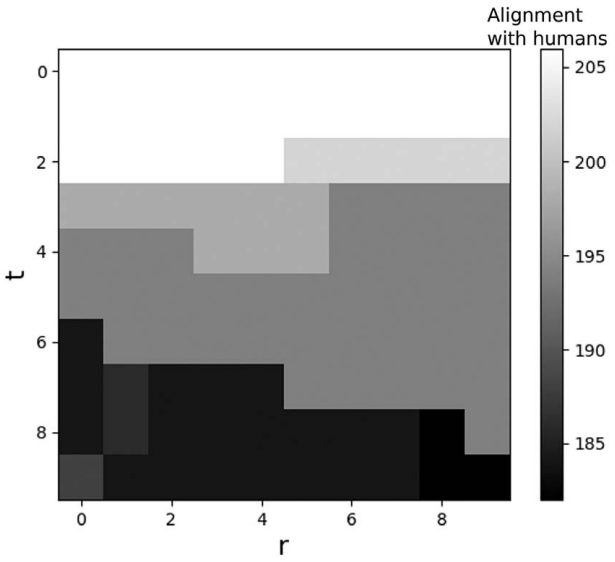

Fig. 3 Plot of performance for different values of order (r) and degree (t) parameters of Sharma–Mittal Entropy. Performance is high for many common entropy metrics like Shannon entropy ($r = 1$, $t = 1$) and HHID ($r = 2$, $t = 2$).

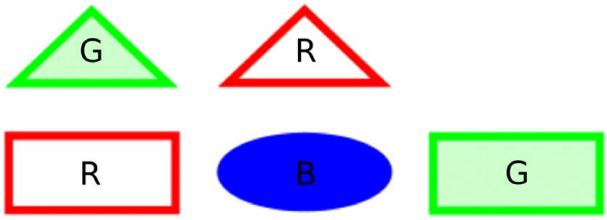

Fig. 4 Set of five polygons with highest variety found using a greedy algorithm applied to submodular objective function

polygon case, the goal is to find a small set of pairwise comparisons of sets, which humans agree on. It is important to create a ground-truth dataset of pairwise queries where human input is most useful in distinguishing between well-known metrics. The process of identifying pairwise comparisons that are shown to human experts is described next.

From the ten sketches, pairwise queries with sets of six sketches have to be created. We decided to create the ground truth with pairs of six sketches as the median number of sketches made by a participant in the entire milk frother dataset [73] was six. The number of unique sets of size six is 210 (10 choose 6). To see the distribution of variety scores and guide the selection of a ground truth dataset, the variety scores for all these sets using SVS and NM metric is calculated. However, in this case, the information about Euclidean embeddings for each sketch (as discussed by Ahmed et al. [75]) is also available, which is used to guide the selection of queries. These embeddings are essentially 2D 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 decide which sets of six sketches to ask humans to rate, information from three metrics (SVS, NM, and average pairwise distance) is used. The last metric is derived using an embedding of designs derived in the study by Ahmed et al. [75]. One design embedding was picked randomly (as each participant in the study had a different design embedding and only one design embedding was needed to guide our experiment), and it provides the 2D positions for each sketch. 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 that are asked from people. The variety scores for all 210 sets was calculated, and all sets were rank ordered from the highest variety set to the lowest variety set, where the variety is measured using the pairwise average distance metric.

Of these 210 sets, we obtained 21,945 pairs of sets (210 choose 2) and calculated the absolute rank difference between the two items for each pair. A small rank difference implies that the two sets in a comparison 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, 20 comparisons were selected based on two factors. First, comparisons were selected where each metric (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, sets with a high-rank difference, but that also differ from sets we are using in other selected comparisons were selected. This ensures that a metric is confident in its vote, but also the queries provide good coverage over different types of sets in the data by ignoring pairs that have already been selected.

Among these candidate sets, 20 pairwise queries are selected that are given to four expert raters using a Qualtrics survey. In the study by Ahmed et al. [75], two participants were considered experts in

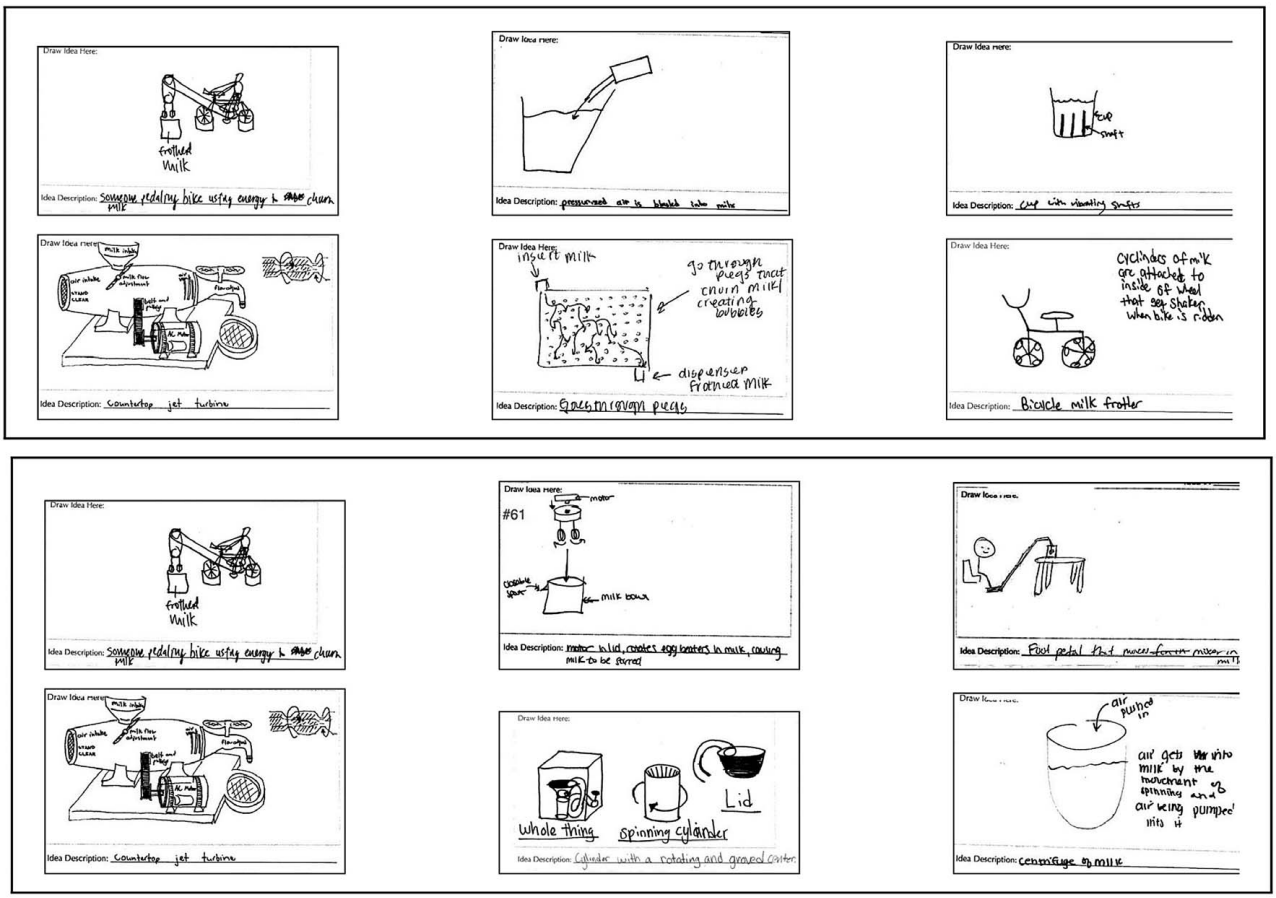

Fig. 5 Top, sample of set A where all raters agreed it was more diverse than set B and bottom, sample of set B where all raters agreed it was less diverse than set A

rating milk frother sketches. To qualify expertise, both experts have at least 4 years of applied experience in design and assessment and had published at least four papers in design and creativity assessment. After analyzing responses on milk frother pairwise comparisons and using similarity metrics, the study also identified that two other participants were clustered together with the two experts and were indistinguishable from experts based on their triplet ratings. We used these four experts for pairwise comparisons in this study. Sets from a typical query is shown Fig. 5. Two comparisons (10% repeated queries) were repeated in each survey to measure the internal consistency of each expert, and a total of 22 queries were given to them. To find the set which covered more design space, experts were asked "Which set of milk frothers has higher variety?" and they can choose whether set A is higher variety compared to set B or they can select the option of "Can't decide." On average, the raters took 24 min to complete the survey. From these expert ratings, we find that all four experts agreed on 9 of 20 queries, while at least three experts agreed on 15 of 20 queries. Due to a majority agreement among experts, these 15 queries are selected as the ground truth dataset for comparing variety metrics. Next, they are used as ground truth dataset to compare variety metrics.

Result 7: SVS and NM Are Equivalent to Random Chance, With Respect to Matching Expert Assessments of Milk Frother Variety. After finding the relative variety scores for each query using SVS and NM, it is seen that they align with only one-third (33.3%) of the ground truth dataset—that is five comparisons. To see if this low performance is due to a specific choice of weights, 1000 possible weight combinations for SVS and NM are tested to 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 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. 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 8: HHID Robustly Outperforms SVS and NM With Respect to Human Comparisons, But Still Has a Nontrivial Error. In contrast to SVS and NM, HHID aligns with 9 of 15 comparisons when weights are optimized for each level. We find that many weight configurations for HHID lead to the highest performance, including $w = [1, 9, 5]$. These weights provide explanations on what factors experts may have given higher priority in their assessments.

Hence, HHID aligns with the human judgment of variety more than both SVS and NM metrics for two standard datasets. However, it still is not 100% aligned to human benchmarks, which can be attributed to the assumption we made, that the annotations provided for SVS, NM, and HHID 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. Constructing the hierarchical trees is outside the scope of this article, 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 aforementioned method for constructing these ground truth variety comparisons, future articles 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.

Discussion

The aforementioned 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 Metrics Is Nontrivial. As we showed earlier, selecting exactly which sets of designs to show experts for ground truth labeling is nontrivial. 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 subsampling 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 [75], 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 overestimate their performance. 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 metrics. 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 optimally (e.g., using Active Learning) and to bound the number of comparisons one would need to collect. Another limitation of our study is that raters may use factors related to the aesthetic aspects of the sketch-like symmetry, the weight of lines, number of pen strokes, amount of white space, handwriting, etc. in deciding which set has a higher variety. These factors, which are unrelated to the design of the milk frother, can influence their decisions in a pairwise comparison, which in turn affects how metrics are judged.

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. Another point to note is the assumptions made in defining the accuracy of a metric. The ground truth defined in this study was based on lots of people agreeing on responses, which is also an indication of repeatability. Ideally, as often observed in the phenomenon studied in physics, the ground truth should be defined as an objectively measurable quantity. However, for a variety metric validation, it is often impossible to have a ground truth that is separate from the collected opinion of others. In such scenarios, it is important to take care of the possible biases introduced in the validation due to the pool of people used to assess the metric.

Metric Performance Can Differ Significantly Across Domains. Comparing Results 4 and Result 7, we see that a given metric applied to one domain/problem may have drastically

different performance. In our case, SVS performed well for 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 detailed earlier. We are releasing both the datasets we collected in this article 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.[7] 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 an objective metric.

Sharma–Mittal entropy Is a Promising Alternative Metric That Allows Optimization of Variety. We demonstrated via Results 4 that using HHID matched or exceed the performance of commonly used metrics. This was true in both the Polygon and milk frother experiments. Calculating the HHID is computationally simpler to the benchmark tree-based constructions of SVS and NM.

More importantly, the submodular form of HHID allows one 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 orders-of-magnitude reductions in computational effort in finding optimal variety subsets of design, compared to existing metrics. The fact that HHID 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.

We further showed in Result 5 that many different settings of Sharma–Mittal entropies are suitable to measure design variety, HHID being one instance of them. We also discussed the conditions under which SME is also submodular, which helps in the optimization of the metric. Different domains tend to use different metrics. This generalization helps one understand why one metric may be more suitable to a particular domain.

It is important to understand a few major assumptions in using objective, numerical metrics. First, SMED is defined for categorical variables (like red, blue, and green), where all categories are assumed to be equally distant from each other. This assumption is used in NM and SVS too. However, it is possible that in some applications, items may have attributes that are real-valued or a few categories can be more similar to each other than others. In such cases, SMED will not be a suitable choice and future work will explore extending SMED to continuous domains. Second, finding the right attributes (or design representation) is critical to the success of any objective design metric. Many manual and automated methods exist to identify suitable attributes for a set of designs. For example, text-based designs may use keyword extraction or topic modeling to identify attributes. Image-based designs may use image descriptors and computer-aided design models may use shape descriptors for attribute identification. This work assumes that the attributes are provided and estimate a variety score for the given set of attributes. Identifying the right attribute to represent different designs is outside the scope of our work.

As both SMED and HHID are derived from entropy metrics, theoretically they can give an absolute score about the variety of a system (in this case, a set of ideas). However, in practice, they are relative metrics. This is because the variety score is dependent on what attributes or categories are considered in the evaluation. If one introduces more categories and reallocates ideas to these new categories, then the variety score may change. For instance, if all "uninteresting" designs are allocated to the same category (given the same attributes), then they will have a small score. However, if one chooses to allocate them different attributes, then the variety score will be large. This limitation, which also exists for other objective metrics, is an artifact of finding the right attributes, and not necessarily of how the metric is defined.

In comparing sets relative to each other, this article also assumes that variety measurements are transitive for a fixed set of attributes (or categories). This assumption is backed by the information–theoretic interpretation of Sharma–Mittal entropy, from which our metric is derived. However, it is possible that human variability judgments are not always transitive, and this assumption was not explicitly tested in this article. As this article uses consensus on pairwise queries to score metrics and not to rank order all sets, this assumption would not affect our results for metric accuracy.

Our future work will focus on using machine learning methods to identify a set of attributes, which are most important in variety estimation. As human judgments are often expensive, an interesting avenue of work will be to cast the fitting of HHID or SME as an active learning problem and reduce the number of expert comparisons needed to adapt design metrics to a new domain. In future work, we will also verify whether human variety judgments are transitive.

Possible Applications of Sharma–Mittal Entropy Beyond Sets. Morphological matrices are a powerful tool for generating ideas based on potential variations in a problem's attributes. For a morphological matrix, the variety score can be calculated in different ways depending on what the end goal is. The morphological matrix is a simple and powerful tool that enables a design engineer to organize and generate all different alternatives before identifying the best design solution [76–78]. SMED can also be applied to morphological matrices. One option is to calculate the variety of the entire matrix, which will inform us how widely do all solutions explore the design space. Another option is to calculate the variety within each function by listing all idea combinations in it and optionally clustering them. In future work, we will explore how variety metrics can be integrated with morphological matrices and compare different ways of doing so.

In applications of morphological matrices and many other design exercises, ideas are often grouped together or chunked during the activity. If ideas are chunked into a set of N_f requirements (or categories/clusters), the SMED score can be calculated using Eq. (3). Hence, the metric allows for the chunking of ideas into groups. If more categories to which an idea can belong are added, it effectively means an increase in the value of N_f in Eq. (3). This means, for the same set of ideas, adding new categories will lead to an increase in the variety score of the set (assuming new categories are not empty), while reducing the number of categories will lead to a reduction in the variety score. In the extreme case, when there is only one category, variety score is zero. Finally, it may also be needed that the variety score is calculated for one set of attributes (or requirements), and later, a different set of attributes is used to calculate the variety score. One can also calculate variety scores using the SMED metric for a different subset of attributes using Eq. (3). In Fig. 2, one may use shape as the only attribute and after the polygons are colored, it may use the color as an attribute to calculate the variety of sets. However, it is important to note that two scores calculated using a different set of attributes cannot be compared with each other meaningfully as the score magnitude also depends on the total number of attributes chosen.

When ideas are collected using methods including brainstorming, the Gallery method, Storyboarding, etc., it is possible that there are missing or incorrectly reported attributes. This situation cannot be

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

handled by existing objective metrics, including SMED. A challenging, albeit an important area of future work is to study design metrics under uncertainty in attribute measurement.

Conclusion

In this article, we contributed (1) a generalization of design variety metric based on the Sharma–Mittal entropy, for which Hirschman–Herfindahl index for design is a special case; (2) a practical procedure for comparing variety metrics via constructing ground truth datasets from pairwise comparisons by experts; and (3) empirically demonstrating the procedure and metric on two new ground truth datasets using milk frother design sketches and polygons. By 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 nontrivial challenges associated with collecting the data needed to measure those qualities. These results guide how and when various commonly used metrics may or may not be valid, as well 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. We hope that by better understanding how to measure the 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 (Grant No. 1728086). We acknowledge the effort of both the MTurk workers and expert raters who help us collect ratings.

Conflicts of Interest

There are no conflicts of interest.

Data Availability Statement

The authors attest that all data for this study are included in the paper. The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request. The data and information that support the findings of this article are freely available at: https://github.com/IDEALLab/design-variety.

References

[1] Amabile, T. M., 1996, *Creativity in Context: Update to the Social Psychology of Creativity*, Hachette, New York.
[2] Sternberg, R. J., 1999, *Handbook of Creativity*, Cambridge University Press, Cambridge, UK.
[3] Mumford, M. D., and Gustafson, S. B., 1988, "Creativity Syndrome: Integration, Application, and Innovation.," Psychol. Bull., **103**(1), p. 27.
[4] Baer, J., 2014, *Creativity and Divergent Thinking: A Task-Specific Approach*, Psychology Press, East Sussex, UK.
[5] Torrance, E. P., 1972, "Predictive Validity of the Torrance Tests of Creative Thinking," J. Creat. Behav., **6**(4), pp. 236–262.
[6] Acar, S., and Runco, M. A., 2017, "Latency Predicts Category Switch in Divergent Thinking.," Psychol. Aesthetics Creat. Arts, **11**(1), p. 43.
[7] Beitz, W., and Pahl, G., 1996, "Engineering Design: A Systematic Approach," MRS Bull., **71**, pp. 63–124.
[8] Shah, J. J., Smith, S. M., and Vargas-Hernandez, N., 2003, "Metrics for Measuring Ideation Effectiveness," Design Stud., **24**(2), pp. 111–134.
[9] Pahl, G., and Beitz, W., 2013, *Engineering Design: A Systematic Approach*, Springer Science & Business Media, New York.
[10] Dorst, K., and Cross, N., 2001, "Creativity in the Design Process: Co-evolution of Problem–Solution," Design Stud., **22**(5), pp. 425–437.
[11] Henderson, D., Helm, K., Jablokow, K., McKilligan, S., Daly, S., and Silk, E., 2017, "A Comparison of Variety Metrics in Engineering Design," ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Cleveland, OH, August.
[12] Maher, M. L., Poon, J., and Boulanger, S., 1996, "Formalising Design Exploration as Co-Evolution," *Advances in Formal Design Methods for CAD*, J. S. Gero, and F. Sudweeks, eds., The International Federation for Information Processing, Springer, Boston, MA, pp. 3–30.
[13] 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," ASME J. Mech. Des., **133**(3), p. 031008.
[14] 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," Res. Engin. Des., **24**(1), pp. 65–92.
[15] Nelson, B. A., Wilson, J. O., Rosen, D., and Yen, J., 2009, "Refined Metrics for Measuring Ideation Effectiveness," Des. Stud., **30**(6), pp. 737–743.
[16] Verhaegen, P.-A., Vandevenne, D., Peeters, J., and Duflou, J. R., 2013, "Refinements to the Variety Metric for Idea Evaluation," Des. Stud., **34**(2), pp. 243–263.
[17] Linsey, J. S., 2007, "Design-by-Analogy and Representation in Innovative Engineering Concept Generation," Ph.D. thesis, University of Texas, Austin, TX.
[18] 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," Des. Stud., **43**, pp. 48–73.
[19] Peeters, J., Verhaegen, P.-A., Vandevenne, D., and Duflou, J., 2010, "Refined Metrics for Measuring Novelty in Ideation," IDMME Virtual Concept Res. Inter. Des., **30**, pp. 20–22.
[20] Fuge, M., Stroud, J., and Agogino, A., 2013, "Automatically Inferring Metrics for Design Creativity," Proceedings of the ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Volume 5: 25th International Conference on Design Theory and Methodology; ASME 2013 Power Transmission and Gearing Conference, Portland, OR, Aug. 4–7, ASME, p. V005T06A010. .
[21] Shatz, D., 2004, *Peer Review: A Critical Inquiry*, Rowman & Littlefield, Lanham, MD.
[22] Chulvi, V., Mulet, E., Chakrabarti, A., López-Mesa, B., and González-Cruz, C., 2012, "Comparison of the Degree of Creativity in the Design Outcomes Using Different Design Methods," J. Engin. Des., **23**(4), pp. 241–269.
[23] Kline, P., 2014, *The New Psychometrics: Science, Psychology and Measurement*, Routledge, London, UK.
[24] Twomey, M., Wallis, L. A., and Myers, J. E., 2007, "Limitations in Validating Emergency Department Triage Scales," Emer. Med. J., **24**(7), pp. 477–479.
[25] Hennessey, B. A., and Amabile, T. M., 1999, "Consensual Assessment," *Encyclopedia of Creativity*, M. Runco, and S. Pritzker, eds., Elsevier Science, San Diego, CA, pp. 347–359.
[26] Harnad, S., 2008, "Validating Research Performance Metrics Against Peer Rankings," Ethics Sci. Environ. Politics, **8**(1), pp. 103–107.
[27] Baer, J., and Kaufman, J. C., 2019, "Assessing Creativity Using the Consensual Assessment Technique," *The Palgrave Handbook of Social Creativity Research. Palgrave Studies in Creativity and Culture*, I. Lebuda and V. Glăveanu, eds., Palgrave Macmillan, Cham, London, UK.
[28] Amabile, T. M., and Pillemer, J., 2012, "Perspectives on the Social Psychology of Creativity," J. Creat. Behav., **46**(1), pp. 3–15.
[29] Schaefer, C. E., and Anastasi, A., 1968, "A Biographical Inventory for Identifying Creativity in Adolescent Boys.," J. Appl. Psychol., **52**(1p1), p. 42.
[30] Taylor, C., and Ellison, R., 1966, *Alpha Biographical Inventory*, Institute for Behavioral Research in Creativity, Salt Lake City, UT.
[31] Cropley, A. J., 2000, "Defining and Measuring Creativity: Are Creativity Tests Worth Using?," Roeper Rev., **23**(2), pp. 72–79.
[32] Douglas, L., Jillian, M., Thomas, L., and Eric, L., 2006, "Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation1," J. Assoc. Infor. Syst., **7**(10), p. 646.
[33] Cross, N., and Roy, R., 1989, *Engineering Design Methods*, Vol. 4, Wiley, New York.
[34] Dylla, N., 1991, "Thinking Methods and Procedures in Mechanical Design," Ph.D. thesis, Mechanical Design, Technical, University of Munich, Munich, Germany.
[35] Song, S., and Agogino, A. M., 2004, "Insights on Designers' Sketching Activities in New Product Design Teams," International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Salt Lake City, UT, September, Vol. 46962, pp. 351–360.
[36] Shah, J. J., 2005, "Identification, Measurement and Development of Design Skills in Engineering Education," DS 35: Proceedings ICED 0n the 15th International Conference on Engineering Design, Melbourne, Australia, Aug. 15–18, pp. 377–378.
[37] Jagtap, S., Larsson, A., Hiort, V., Olander, E., and Warell, A., 2015, "Interdependency Between Average Novelty, Individual Average Novelty, and Variety," Int. J. Des. Creat. Innovat., **3**(1), pp. 43–60.
[38] Jansson, D. G., and Smith, S. M., 1991, "Design Fixation," Des. Stud., **12**(1), pp. 3–11.

[39] Kershaw, T. C., and Ohlsson, S., 2004, "Multiple Causes of Difficulty in Insight: The Case of the Nine-Dot Problem.," J. Exp. Psychol. Learn. Memory Cognition, **30**(1), p. 3.
[40] Bonnardel, N., 2000, "Towards Understanding and Supporting Creativity in Design: Analogies in a Constrained Cognitive Environment," Knowl. Based Syst., **13**(7–8), pp. 505–513.
[41] Herring, S. R., Chang, C. -C., Krantzler, J., and Bailey, B. P., 2009, "Getting Inspired! Understanding How and Why Examples Are Used in Creative Design Practice," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, April, pp. 87–96.
[42] Wilson, J. O., Rosen, D., Nelson, B. A., and Yen, J., 2010, "The Effects of Biological Examples in Idea Generation," Design Stud., **31**(2), pp. 169–186.
[43] Srinivasan, V., and Chakrabarti, A., 2010, "Investigating Novelty–Outcome Relationships in Engineering Design," AI EDAM, **24**(2), pp. 161–178.
[44] Verhaegen, P.-A., Vandevenne, D., Peeters, J., and Duflou, J. R., 2015, "A Variety Metric Accounting for Unbalanced Idea Space Distributions," Procedia. Eng., **131**, pp. 175–183.
[45] Viswanathan, V. K., and Linsey, J. S., 2012, "Physical Models and Design Thinking: A Study of Functionality, Novelty and Variety of Ideas," ASME J. Mech. Des., **134**(9), p. 091004.
[46] Thevenot, H. J., and Simpson, T. W., 2007, "A Comprehensive Metric for Evaluating Component Commonality in a Product Family," J. Engin. Des., **18**(6), pp. 577–598.
[47] Kota, S., Sethuraman, K., and Miller, R., 2000, "A Metric for Evaluating Design Commonality in Product Families," ASME J. Mech. Des., **122**(4), pp. 403–410.
[48] Jung, S., and Simpson, T. W., 2016, "An Integrated Approach to Product Family Redesign Using Commonality and Variety Metrics," Res. Engin. Des., **27**(4), pp. 391–412.
[49] Chan, J., Dang, S., and Dow, S. P., 2016, "Comparing Different Sensemaking Approaches for Large-Scale Ideation," Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, May, ACM, pp. 2717–2728.
[50] 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," ASME J. Mech. Des., **133**(8), p. 081004.
[51] Gini, C., 1912, "Variabilità E Mutabilità," *Reprinted in Memorie Di Metodologica Statistica*, E. Pizetti, and T. Salvemini, eds., Libreria Eredi Virgilio Veschi, Rome.
[52] Simpson, E. H., 1949, "Measurement of Diversity," Nature, **163**(4148), p. 688.
[53] Shannon, C. E., 1948, "A Mathematical Theory of Communication," Bell. Syst. Tech. J., **27**(3), pp. 379–423.
[54] Jost, L., 2006, "Entropy and Diversity," Oikos, **113**(2), pp. 363–375.
[55] Hirschman, A. O., 1964, "The Paternity of an Index," Amer. Econom. Rev., **54**(5), pp. 761–762.
[56] Crupi, V., 2019, "Measures of Biological Diversity: Overview and Unified Framework," *From Assessing to Conserving Biodiversity. History, Philosophy and Theory of the Life Sciences*, E. Casetta, J. Marques da Silva, and D. Vecchi, eds., vol. 24, Springer, Cham.
[57] Rhoades, S. A., 1993, "The Herfindahl-Hirschman Index," Fed. Res. Bull., **79**, p. 188.
[58] Greenberg, J. H., 1956, "The Measurement of Linguistic Diversity," Language, **32**(1), pp. 109–115.
[59] Pitt, R., Pirtle, W. N. L., and Metzger, A. N., 2019, "Academic Specialization, Double Majoring, and the Threat to Breadth in Academic Knowledge," J. General Edu., **66**(3-4), pp. 166–191.
[60] Nelson, J. D., Crupi, V., Meder, B., Cevolani, G., and Tentori, K., 2017, "A Unified Model of Entropy and the Value of Information," Decision Making, **7**, pp. 119–148.
[61] Masi, M., 2005, "A Step Beyond Tsallis and Rényi Entropies," Phys. Lett. A., **338**(3–5), pp. 217–224.
[62] Ahmed, F., Ramachandran, S. K., Fuge, M., Hunter, S., and Miller, S., 2019, "Measuring and Optimizing Design Variety Using Herfindahl Index," Proceedings of the ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Volume 7: 31st International Conference on Design Theory and Methodology, Anaheim, CA, Aug. 18–21, ASME, p. V007T06A007.
[63] Hurlbert, S. H., 1971, "The Nonconcept of Species Diversity: A Critique and Alternative Parameters," Ecology, **52**(4), pp. 577–586.
[64] Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L., 1978, "An Analysis of Approximations for Maximizing Submodular Set Functionsi," Math. Program., **14**(1), pp. 265–294.
[65] Feige, U., Mirrokni, V. S., and Vondrak, J., 2011, "Maximizing Non-Monotone Submodular Functions," SIAM J. Comput., **40**(4), pp. 1133–1153.
[66] Krause, A., and Golovin, D., 2014, "Submodular Function Maximization," *Tractability: Practical Approaches to Hard Problems*, L. Bordeaux, Y. Hamadi, and P. Kohli, eds., Cambridge University Press, Cambridge, UK, pp. 71–104.
[67] Hoffmann, S., 2008, "Generalized Distribution Based Diversity Measurement: Survey and Unification," Faculty of Economics and Management, Otto-von-Guericke University Magdeburg, Technical Report 08023.
[68] Stobbe, P., and Krause, A., 2010, "Efficient Minimization of Decomposable Submodular Functions," Advances in Neural Information Processing Systems, Vancouver, Canada, December, pp. 2208–2216.
[69] Kendall, M., 1962, *Rank Correlation Methods* (Theory and Applications of Rank Order-Statistics), Hafner Pub. Co, New York City, NY.
[70] Stewart, N., Brown, G. D., and Chater, N., 2005, "Absolute Identification by Relative Judgment.," Psychol. Rev., **112**(4), p. 881.
[71] Ahmed, F., Fuge, M., and Gorbunov, L. D., 2016, "Discovering Diverse, High Quality Design Ideas From a Large Corpus," Proceedings of the ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Volume 7: 28th International Conference on Design Theory and Methodology, Charlotte, NC, Aug. 21–24, ASME, p. V007T06A008.
[72] Ahmed, F., and Fuge, M., 2018, "Ranking Ideas for Diversity and Quality," ASME J. Mech. Des., **140**(1), p. 011101.
[73] 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," ASME J. Mech. Des., **141**(1), p. 012001.
[74] Toh, C. A., and Miller, S. R., 2014, "The Impact of Example Modality and Physical Interactions on Design Creativity," ASME J. Mech. Des., **136**(9), p. 091004.
[75] Ahmed, F., Ramachandran, S. K., Fuge, M., Hunter, S., and Miller, S., 2019, "Interpreting Idea Maps: Pairwise Comparisons Reveal What Makes Ideas Novel," ASME J. Mech. Des., **141**(2), p. 021102.
[76] Ritchey, T., 2006, "Problem Structuring Using Computer-Aided Morphological Analysis," J. Oper. Res. Soc., **57**(7), pp. 792–801.
[77] Pahl, G., Beitz, W., Feldhusen, J., and Grote, K., 2007, *Engineering Design: A Systematic Approach* (Solid Mechanics and Its Applications), Springer, London.
[78] George, D., Linnerud, B., and Mocko, G., 2014, "Integrated Idea Generation Method for Concept Generation Using Morphological and Options Matrices," Proc. TMCE, Budapest, Hungary, May, pp. 1–12.