Context Matters: A Theory of Semantic Discriminability for Perceptual Encoding Systems

Kushin Mukherjee, Brian Yin, Brianne E. Sherman, Laurent Lessard, and Karen B. Schloss

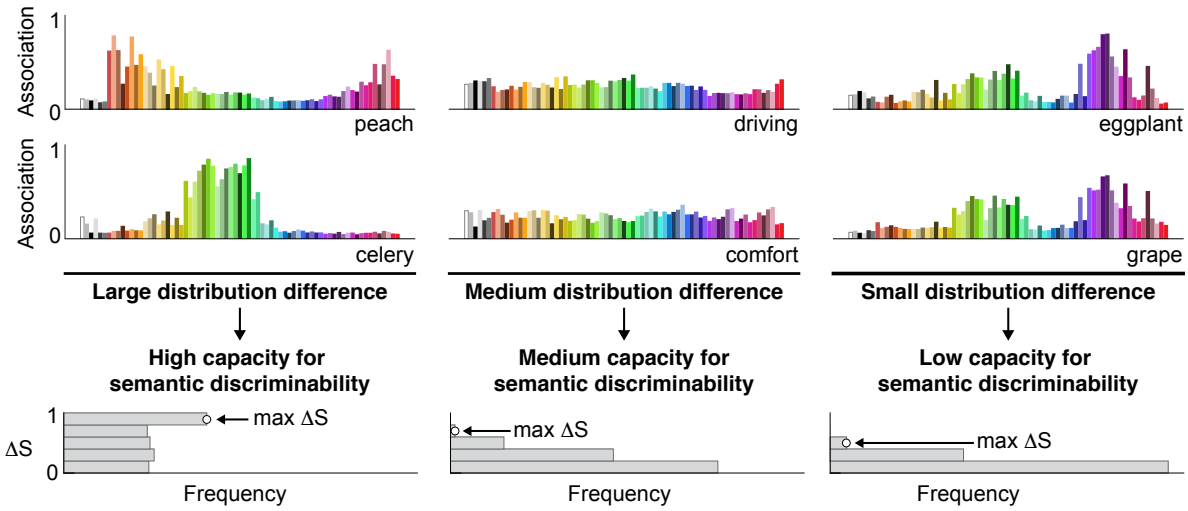

Figure 1: Color-concept association distributions for concept pairs with large, medium, and small distribution differences, resulting in high, medium, and low capacities for semantic discriminability, respectively (terms defined in Section 3). Color-concept association ratings were collected in Experiment 1 for the UW-71 colors (colored stripes in the plots, sorted by CIE LCh hue angle).

Abstract— People's associations between colors and concepts influence their ability to interpret the meanings of colors in information visualizations. Previous work has suggested such effects are limited to concepts that have strong, specific associations with colors. However, although a concept may not be strongly associated with any colors, its mapping can be disambiguated in the context of other concepts in an encoding system. We articulate this view in semantic discriminability theory, a general framework for understanding conditions determining when people can infer meaning from perceptual features. Semantic discriminability is the degree to which observers can infer a unique mapping between visual features and concepts. Semantic discriminability theory posits that the capacity for semantic discriminability for a set of concepts is constrained by the difference between the feature-concept association distributions across the concepts in the set. We define formal properties of this theory and test its implications in two experiments. The results show that the capacity to produce semantically discriminable colors for sets of concepts was indeed constrained by the statistical distance between color-concept association distributions (Experiment 1). Moreover, people could interpret meanings of colors in bar graphs insofar as the colors were semantically discriminable, even for concepts previously considered "non-colorable" (Experiment 2). The results suggest that colors are more robust for visual communication than previously thought.

Index Terms—Visual Reasoning, Information Visualization, Visual Communication, Visual Encoding, Color Cognition

1 Introduction

Bananas are shades of yellow, blueberries are shades of blue, and cantaloupes are shades of orange. It is well-established that color semantics influences people's ability to interpret information visualizations when those visualizations represent concepts that have specific, strongly associated colors (e.g., fruits). Such visualizations are easier to interpret if concepts are encoded with strongly associated colors (e.g., bananas encoded with yellow, not blue) [20, 31]. But, how often do real-world visualizations really depict information about fruit, or other concepts with specific, strongly associated colors? If color semantics mainly influences interpretability for visualizations of concepts with specific, strongly associated colors (as previously suggested [20, 33]), then scenarios in which color semantics matters would be severely limited.

The present study suggests people's ability to infer meaning from colors is more robust than previously thought. Conditions arise in which people can interpret meanings of colors for concepts previously considered "non-colorable". Specifically this when the colors are semantically discriminable. Semantic discriminability for colors is the ability to infer unique mappings between colors and concepts based on colors and concepts alone (i.e., without using a legend) [31]. This is distinct from *semantic interpretability*, which is the ability to interpret the *correct* mapping between colors and concepts, as specified in an encoding system (for further discussion of this distinction, see [31] and Supplementary Material Section S.7 in the present paper). The key

- *Kushin Mukherjee, Psychology and Wisconsin Institute for Discovery, University of Wisconsin–Madison . Email: kmukherjee2@wisc.edu.*
- *Brian Yin, Cognitive Science, University of California, Berkeley. Email:brianyin@berkeley.edu.*
- *Brianne E. Sherman, Neurobiology and Wisconsin Institute for Discovery, University of Wisconsin–Madison, Email: besherman2@wisc.edu.*
- *Laurent Lessard, Mechanical and Industrial Engineering, Northeastern University. Email: l.lessard@northeastern.edu.*
- *Karen B. Schloss, Psychology and Wisconsin Institute for Discovery, University of Wisconsin–Madison. Email: kschloss@wisc.edu.*

Manuscript received 21 Mar. 2021; revised 13 June 2021; accepted 8 Aug. 2021. Date of publication 29 Sept. 2021; date of current version 22 Dec. 2021. Digital Object Identifier no. 10.1109/TVCG.2021.3114780

question is, what determines whether it is possible to select semantically discriminable colors for a set of concepts?

We address this question in *semantic discriminability theory*, a new theory on constraints for generating semantically discriminable perceptual features for encoding systems that map perceptual features to concepts. We tested two hypotheses that arise from the theory. First, the capacity to create semantically discriminable color palettes for a set of concepts depends on the difference in color-concept association distributions *between* those concepts, independent of properties of the concepts in isolation (Experiment 1). Second, people can accurately interpret mappings between colors and concepts for concepts previously considered "non-colorable," to the extent that the colors are semantically discriminable (Experiment 2). We focus on color in this study, but present the theory in terms of perceptual features more generally because of its potential to extend to other types of visual features (e.g., shape, orientation, visual texture) and features in other perceptual modalities (e.g., sound, odor, touch).

Contributions. This paper makes the following contributions: (1) We define semantic discriminability theory (Section 3) and test hypotheses motivated by the theory in Experiments 1 and 2 (Sections 4-5), and (2) We define a new metric for operationalizing distribution difference between sets of more than two concepts (Section 3.2) and show that it predicts capacity for semantic discriminability (Section 4).

2 Background

Color is a strong cue for signaling meaning in nature and some argue that color vision evolved for the purpose of visual communication [7,8,12,14,39]. Historically, discussions on the role of color semantics in information visualization have tended to focus on few cases of typical associations (e.g., red for hot, green for grass) [5, 28, 34]. More recent work has sought to understand the potential and limitations of using color to communicate meaning in visualizations [1, 2, 20, 31–33]. The semantics of color in visualizations operates on two main levels: meaning of a color palette as a whole [1, 2, 15] and meaning of the individual colors in a palette [20, 31–33]. We focus on meanings of individual colors because that is central to the present work. People have expectations about how colors will map onto concepts, and visualizations that violate those expectations are harder to interpret, even if there is a legend [20, 30, 35]. Thus, understanding these expectations is important for optimizing palette design for visual communication.

2.1 Color-concept associations

Color-concept associations represent the degree to which individual colors are associated with individual concepts. Color-concept associations can be quantified using various methods, including human judgments [16, 17, 25, 27, 31, 32, 38], image statistics [20–22, 27, 33], and natural language corpora [13, 33]. Some approaches focused on identifying the strongest, or strongest few colors associated with a concept [11, 13, 33], but color-concept associations can be treated as a continuous property over all possible colors in a perceptual color space [20–22, 27, 29]. When quantifying color-concept associations over all of color space, researchers typically bin or sub-sample parts of the space to make measurements computationally tractable. An assumption is that the space is continuous, so nearby colors will have similar associations. Figure 1 shows examples of color-concept associations for colors systematically sampled over CIELAB space (see Experiment 1), plotted over one dimension (sorted by hue angle and chroma with achromatics at the beginning of the list). Perceptual color spaces are three-dimensional so this representation does not necessarily position perceptually similar colors in close proximity [41], but it does highlight how some concepts, like peach and celery, have specific, strongly associated colors, whereas other concepts, like driving and comfort, are more uniform (Figure 1). We refer to this 'peakiness' property as *specificity* of the color-concept association distributions.[1]

[1]Specificity is similar to color diagnosticity [37], but color diagnosticity concerns whether a concept has a single strongly associated color [37], and specificity concerns the degree to which a concept is associated with some colors more than others in a color-concept association distribution.

Questions remain concerning how color concept-associations are formed, but many have suggested that they are learned through experiences [10, 16, 27, 38, 40] and may be continually updated through each new experience in the world [29]. Some color-concept associations are shared cross-culturally, and others are subject to cultural differences [16, 17, 38]. We will consider the role of cultural differences with respect to the present work in the General Discussion.

Color-concept associations contribute to people's expectations about the meanings of colors in information visualizations [20, 31, 32], called *inferred mappings*. However, associations and inferred mappings are not the same, and sometimes they conflict [32]. We explain this point in Section 2.3 on assignment inference.

2.2 Colorabilty scores

Some have suggested that the effectiveness of colors for encoding meaning is limited to concepts that have strong associations with particular colors [18, 20, 33]. This idea is explained by invoking *colorability* scores, which broadly measure how strongly *individual* concepts can be mapped to specific colors. Generally, concepts with specific, strongly associated colors ('banana') are thought to be colorable, whereas more abstract concepts, such as 'comfort' or 'leisure', that lack such strongly associated colors, have been called non-colorable.

Different methods have been used to define colorability. Lin et al. [20] quantified colorability by having participants assign colors to concepts and rate the strength of the assignment. The mean of these ratings over all colors for a concept was used to generate a colorability score for that concept. They found that participants were better at interpreting bar charts when palettes were optimized for color semantics compared to when palettes had the default Tableau color ordering, but this benefit was mostly limited to highly colorable concepts. Setlur and Stone [33] quantified colorability with an automated method, using Google N-grams to determine how frequently a concept word co-occurred with basic color terms [3] in linguistic corpora. They then excluded concepts they found to be non-colorable when developing methods to optimize palette design.

These prior studies highlighted the importance of considering color semantics in palette design. However, our work suggests that restricting notions of colorability to concepts in isolation may have led to underestimating people's ability to infer meaning from colors in visualizations.

2.3 Assignment inference

Evidence suggests that people's inferences about the meanings of colors in encoding systems of visualizations do not merely depend on color-concept associations in isolation. We illustrate this point with an example from Schloss et al. [32]. Participants saw pairs of unlabeled bins and were asked to choose which bin was for the target concept written at the top of the screen. Figure 2 shows two examples when trash was the target concept. The other concept, not pictured here but judged on other trials, was paper. To the left of the example trials are bipartite graphs, which use line thickness to represent the association strength between each concept (trash, T, and paper, P) and each color in the corresponding trial. An easy way to approach this task would be to choose the color that is most strongly associated with trash within each trial (local assignment). Alternatively, participants could choose the color that results in maximizing association strengths of all color-concept pairings across trials (global assignment).

In the top row of Figure 2, these two approaches lead to the same outcome. Locally, trash is more strongly associated with dark yellow (Y) than white (W). Globally, the assignment trash-yellow/paper-white has a larger overall association strength than trash-white/paper-yellow. Not surprisingly, participants inferred trash is mapped to dark yellow. However, in the bottom row, the two approaches lead to opposite outcomes. Locally, trash is more associated with white than purple (Pu), but globally the assignment trash-purple/paper-white has a larger overall association strength (greater total thickness of edges) than trash-white/paper-purple. Participants inferred that trash maps to purple, even though white was a more strongly associated alternative. Each trial was independent, so participants need not account for paper on trials for trash, but they did so nonetheless. This example highlights the

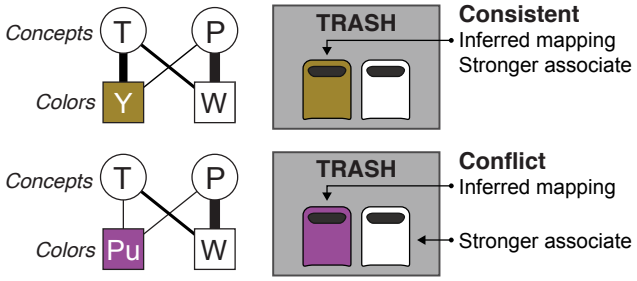

Figure 2: Distinction between color-concept associations and inferred mappings (figure based on [32]). Left: Bipartite graphs show color-concept association strengths for concepts trash (T) and paper (P) with colors dark yellow (Y), white (W), and purple (Pu) (thicker edges connecting concepts and colors indicate stronger associations). Right: example trials where participants infer which color maps to trash.

important distinction between color-concept associations for a single color and concept, and inferred mappings between a color and concept in the context of an encoding system.

Schloss et al. [32] called this process of inferring mappings between colors and concepts *assignment inference* because it is analogous to an *assignment problem* in optimization. In assignment problems, every possible pairing of items in one category (e.g., colors) i and another category (e.g., concepts) j is given a numerical *merit score* m_{ij}. Here, let's assume that larger scores indicate a more desirable pairing, but that is not always true (e.g., to optimize delivery route efficiency, merit might be delivery time and smaller scores would be better). Solving an assignment problem means finding the pairing of items that maximizes (or minimizes) the sum of the merit scores of all chosen pairs [6, 19, 23].

Although assignment inference is analogous to assignment problems, they are not the same. Assignment problems have deterministic results, whereas assignment inference is stochastic—inferred mappings can vary among individuals and even within individuals over time. This stochasticity can be explained in terms of noise in people's color-concept associations affecting the outcome of assignments in assignment inference, depending on whether assignments are robust or fragile [31]. In robust assignments, adding noise to the system (e.g., perturbing the color-concept association strengths) has no effect on the outcome, but in fragile assignments adding noise can change the outcome of the assignment.

The robustness of an assignment in assignment inference can be understood as *semantic discriminability*—the ability for people to infer a unique mapping between colors and concepts [31]. Evidence suggests that semantic discriminability predicts people's ability to interpret colors in encoding systems, independent of that predicted by perceptual discriminability and color-concept associations in isolation [31]. We describe ways of operationalizing semantic discriminability in Section 3.2 as they pertain to the present study.

So far, we focused on encoding systems with two concepts and colors, and implied that merit m_{ij} in assignment inference is color-concept association strength (Figure 2). However, there are other possible ways to define merit, especially when there are more than two colors and concepts, as in the present study. Schloss et al. [32] sought to understand which merit people use in assignment inference to study (1) how humans infer meaning from colors, and (2) how to design palettes that match people's expectations, making palettes more interpretable. To approach this goal, they created two definitions of merit. The *isolated merit function* simply uses association strengths between items i and j, $m_{ij} := a_{ij}$. The *balanced merit function* is defined as

$$m_{ij} := a_{ij} - \max_{k \neq j} a_{ik}. \tag{1}$$

The balanced merit score for a given color-concept pair is the association strength for that pair, minus the association strength between that color and the next most strongly associated concept. In order for m_{ij} to be large, color i should be strongly associated with concept j and weakly associated with all other concepts. (Note: in the case of two concepts and colors these two definitions reduce to the same outcome.)

Next, they generated color palettes using an assignment problem under each definition, with human color-concept association ratings as the input. Finally, they presented different participants with those palettes in the form of six unlabeled colored bins. Participants inferred which bin was for each of six objects: paper, plastic, trash, metal, compost, and glass. Responses were scored as "correct" interpretations if they matched the encoded mapping. Encoded mappings can be produced in different ways, including by designers, software defaults, or optimization algorithms [20, 31, 32]. Here, they were determined by the optimal assignments in assignment problems used to generate the palettes. The logic was that participants would be better at interpreting palettes generated using a merit function that more closely matched merit in assignment inference. Performance was better for the palette generated using the balanced merit function, which suggests that this was the function that better captured merit in assignment inference. Thus, we use balanced merit in the present study.

Balanced merit can lead to unexpected assignments. For example, the bin for plastic was assigned a red color, even though red was weakly associated with plastic, because that color was more associated with plastic than with any of the other concepts. Thus, the assignment of plastic–red was interpretable. Given that weakly associated colors can prove useful when designing encoding systems, approaches that focus only on the top associates may be limited [11]. It is important to quantify associations between concepts and a large range of colors, not just the top few associates, when optimizing palette design [27].

3 Semantic Discriminability Theory

Semantic discriminability theory characterizes the ability to generate semantically discriminable perceptual features for encoding a set of concepts. We begin with some key definitions.

Concept set: This is the set of all concepts that are represented in an encoding system. These concepts could refer to any information that is categorical (e.g., food, weather, activities, places, and animals). We label concepts in the concept set using the index $j \in \{1, 2, \ldots, n\}$.

Feature source: This is the set of all possible instances of a feature type. Perceptual color spaces (e.g., CIELAB) are well-defined feature sources for color, as they represent all colors humans can perceive [41].

Feature library: This is a subset of candidate features from the feature source used in an encoding system. For example, the Tableau 20 colors or UW-58 colors [27] are feature libraries if design is constrained to those groups of colors. We focus on a feature library defined over color, but they can be defined over any type of perceptual feature (e.g., shapes, sizes, textures). We label features in the feature library using the index $i \in \{1, 2, \ldots, N\}$.

Feature set: This is a subset of features from the feature library, selected to encode a concept set. Feature sets can be constructed from any type of perceptual features (e.g., colors, shapes, sizes) [4]. For colors, they are called "palettes." If there are n concepts, then the feature set should contain n features.

3.1 Feature-concept association distributions

Feature-concept association distributions represent the degree to which a given concept is associated with each feature in a feature library (see Figure S.5A. in the Supplementary Material). For color, these are color-concept association distributions. Feature-concept association distributions can be described as raw association values over the feature library (e.g., mean ratings, pixel counts, word counts). In this case, we write a_{ij} to denote the association between feature $i \in \{1, \ldots, N\}$ and concept $j \in \{1, \ldots, n\}$. For each concept j, we also define *normalized* associations $p_j(\cdot)$ as

$$p_j(i) := \frac{a_{ij}}{\sum_{k=1}^{N} a_{kj}} \qquad \text{for: } i \in \{1, \ldots, N\}. \tag{2}$$

The list $\begin{bmatrix} p_j(1) & p_j(2) & \cdots & p_j(N) \end{bmatrix}$ can be interpreted as a discrete probability distribution over features in the feature library.

We now define useful properties and operations related to feature-concept association distributions.

3.1.1 Specificity

Specificity is the degree to which a concept has strong, specific associations with features over the feature library. For color, specificity refers to the 'peakiness' of a color-concept association distribution. Concepts can have strong color associations that are concentrated in one part of color space (e.g., reds for concepts like raspberry) or divided over different parts of color space (e.g., reds and greens for watermelon) [27]. Thus, we quantify specificity using *entropy* of the distribution, which captures how 'flat' vs. 'peaky' a distribution is, regardless of how many peaks there are.

Entropy for a feature-concept association distribution is defined as:

$$H_j := -\sum_{i=1}^{N} p_j(i) \log p_j(i). \tag{3}$$

If all features in the feature library are equally associated with concept j, the distribution p_j will be uniform, entropy will be high, and specificity will be low. If a concept j is strongly associated with some features and not others, then entropy will be lower and specificity will be higher. This property of color-concept association distributions aligns with previous measures of colorability [20, 33] (see Figure S.2 in the Supplementary Material).

Mean entropy of a concept set is the mean of the entropy of all concepts in the set: $H_\mu := \frac{1}{n}(H_1 + \cdots + H_n)$.

3.1.2 Distribution difference

We quantify distribution difference between concepts by comparing their normalized feature-concept associations.

Total variation (TV) is what we use when comparing two concepts, say j_1 and j_2. TV is defined as follows.

$$\mathrm{TV}(j_1, j_2) := \frac{1}{2} \sum_{i=1}^{N} \left| p_{j_1}(i) - p_{j_2}(i) \right|. \tag{4}$$

TV ranges between 0 and 1, where $\mathrm{TV} = 0$ means the two distributions are identical, and $\mathrm{TV} = 1$ means they are disjoint (for each feature i, either $p_{j_1}(i)$ or $p_{j_2}(i)$ must be zero).

Generalized total variation (GTV) is a generalization of TV that we defined for cases when more than two concepts must be compared, say $j_1, \ldots, j_k$. We define GTV as follows.

$$\mathrm{GTV}(j_1, \ldots, j_k) := -1 + \sum_{i=1}^{N} \max(p_{j_1}(i), p_{j_2}(i), \ldots, p_{j_k}(i)). \tag{5}$$

In the case where $k = 2$, GTV reduces to TV. In other words, $\mathrm{GTV}(j_1, j_2) = \mathrm{TV}(j_1, j_2)$. For details on the motivation behind our definition of GTV, see the Supplementary Material, Section S.6.

3.1.3 Structure-agnostic property

The notions of entropy, TV, and GTV are agnostic to intrinsic structure of the feature source. For example, perceptual color spaces are structured according to perceptual similarity, but entropy of a color-concept distribution depends on the fraction of the colors that are highly associated with the concept, regardless of perceptual similarity. We chose structure-agnostic metrics for specificity and distribution difference so that semantic discriminability theory could readily generalize to feature sources with less well-defined metric spaces (e.g., shape, texture, odor).

3.2 Semantic discriminability

As described in Section 2.3, semantic discriminability of perceptual features is the ability to infer a unique mapping between features and concepts. It is reflected in the degree to which inferred mappings vary among individuals or within individuals between trials. We model this variability by treating feature-concept associations as random variables. Rather than solving an assignment problem using the mean a_{ij} values, we look at the *probability* of the likeliest assignment, where probability is computed with respect to uncertainty in the a_{ij}. We now make this notion more precise.

Semantic distance is a way to operationalize semantic discriminability in the case where there are $n = 2$ features and concepts [31]. Figure 3 illustrates an example in which we have concepts {M,W} and colors {1,2}. The color-concept associations between all possible pairs are $x_1, \ldots, x_4$, as shown in Figure 3. We assume each x_k is normally distributed with mean $\bar{x}_k$ equal to the corresponding a_{ij} and standard deviation $\sigma_k = 1.4 \cdot \bar{x}_k(1 - \bar{x}_k)$, which was found to be a good fit to experimental data [31]. The outcome of the assignment problem is determined by the quantity $\Delta x := x_1 - x_2 + x_3 - x_4$. The optimal assignment is: (M-1 and W-2 if $\Delta x > 0$) and (M-2 and W-1 if $\Delta x < 0$). Semantic distance is defined by the equation

$$\Delta S = |\mathrm{Prob}(\Delta x > 0) - \mathrm{Prob}(\Delta x < 0)|. \tag{6}$$

Since the x_k are assumed to be normally distributed, so is Δx, and the probabilities in (6) can be computed analytically:

$$\mathrm{Prob}(\Delta x > 0) = \Phi\left(\frac{(\bar{x}_1 + \bar{x}_4) - (\bar{x}_2 + \bar{x}_3)}{\sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2}} \right), \tag{7}$$

and $\mathrm{Prob}(\Delta x < 0) = 1 - \mathrm{Prob}(\Delta x > 0)$, where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal distribution. When ΔS is close to 0, Δx has a similar probability of being positive or negative, so the assignment is fragile. When ΔS is close to 1, Δx is almost always positive or almost always negative, so the assignment is robust. This notion of semantic distance can be used even when the features are not colors, by replacing the color-concept associations with feature-concept associations, and adjusting the formula for σ_k as appropriate.

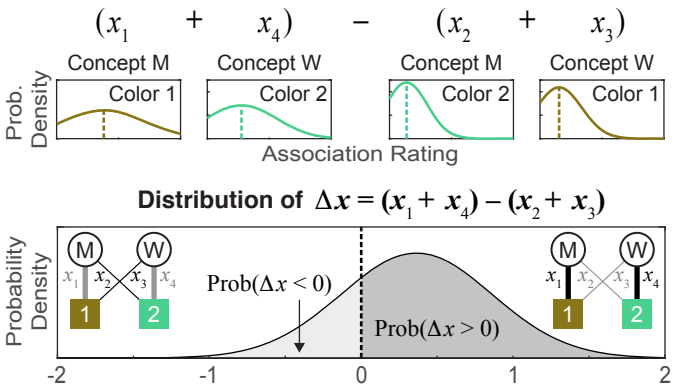

Figure 3: Diagram from [31] that shows how association ratings between concepts {M,W} and colors {1,2} produce a distribution for Δx. Semantic distance is the absolute difference of the area under the curve to the left and right of zero.

Generalized semantic distance is an extension of semantic distance to the case where there are $n > 2$ features and concepts. In this case, there will be $n!$ (n factorial) possible assignments. We define generalized semantic distance in a manner analogous to semantic distance; we label the feature-concept associations between all possible pairs as $x_1, x_2, \ldots, x_{n^2}$ and assume they are normally distributed random variables.[2] In this more complicated scenario, the assignment is not determined by a simple quantity such as Δx and no formula analogous to (7) exists to determine the assignment. Instead, we use the following Monte Carlo approach.

1. Sample $x_1, \ldots, x_{n^2}$ from the distribution of merit scores[2] and solve an assignment problem using the sampled merit scores.
2. Repeat step 1 a large number of times and count the number of times each distinct assignment occurs. Let p be the proportion of times that the most frequent assignment occurred. Since there are $n!$ possible assignments, we must have $\frac{1}{n!} \leq p \leq 1$.

[2]Here, we use color-concept association ratings, so we assume the x_k are distributed with the same σ_k used to define semantic distance [31]. In principle, the distributions of the x_k can be changed to suit other cases beyond color.

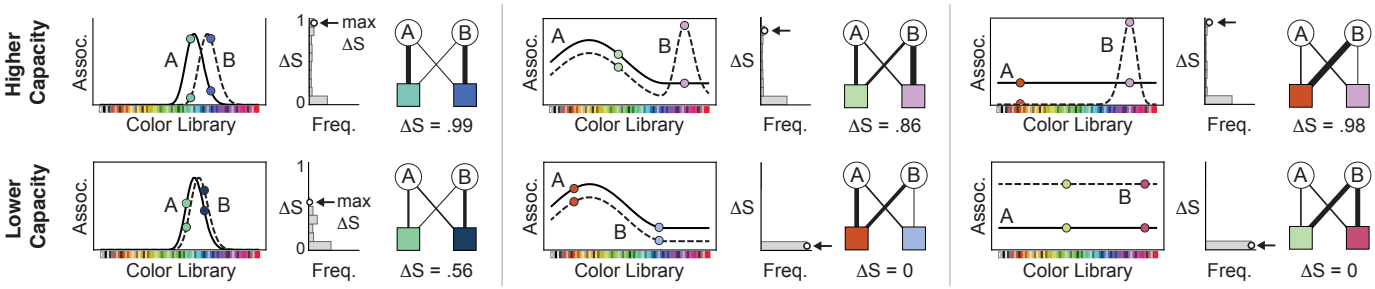

Figure 4: Hypothetical color-concept association distributions for concepts A and B, showing how capacity varies with distribution difference (top row: higher capacity; bottom row: lower capacity). In each column, the distribution for concept A is the same and concept B varies. The histograms to the right show how distribution difference affects capacity with arrows pointing to maximum semantic distance (ΔS) for the concept set. Corresponding bipartite graphs show the color set with maximum semantic distance (this is arbitrary when the distributions are parallel because semantic distances for all color pairs are equally poor).

3. The generalized semantic distance ΔS is defined as a linear rescaling of p to ensure that $0 \leq \Delta S \leq 1$. The formula is:

$$\Delta S = \frac{n!p - 1}{n! - 1}. \tag{8}$$

A similar Monte Carlo approach was used in [32] to predict the results of assignment inference in a recycling task (6 concepts and 6 colors).

Just like semantic distance, generalized semantic distance is a number between 0 and 1, where a larger number indicates more robust assignments, and consequently, higher semantic discriminability. We use the same symbol ΔS for both notions of distance because in the case where $n = 2$, generalized semantic distance is (on average) equal to semantic distance, and the approximation becomes exact as the number of samples in step 2 tends to infinity. Conversely, in the limit $n \to \infty$, we have $\Delta S \to p$ and the rescaling in (8) has no effect.

Semantic contrast is similar to generalized semantic distance, except it estimates the proportion of times a given color is assigned to the "optimal" concept (compared to all other assignments). This estimation is computed using the Monte Carlo method described earlier, with optimal defined by the solution to an assignment problem using the balanced merit function computed on feature-concept associations.

For a given concept, the optimal color for that concept may have higher semantic contrast in one context and lower semantic contrast in another context, depending on the other colors and concepts in the encoding system. A concept set that has higher capacity for semantic discriminability (Section 3.3) should enable higher semantic contrasts among colors in its optimal palette.

The steps to computing semantic contrast are: (1) Solve an assignment problem (see Section 2.3) using the mean association ratings $\bar{x}_1, \ldots, \bar{x}_{n^2}$. We call this the *optimal assignment*. (2) Sample $x_1, \ldots, x_{n^2}$ from the distribution of merit scores and solve an assignment problem using the sampled merit scores. (3) Repeat step 2 a large number of times and count the proportion of times each feature was assigned to the same concept as in the optimal assignment This proportion is each feature's semantic contrast.

3.3 Capacity for semantic discriminability

Capacity for semantic discriminability is the extent to which it is possible to produce semantically discriminable features for a given set of concepts. We operationalized capacity for semantic discriminability (*capacity* for short), using **max capacity**. This is a scalable measure that returns the semantic distance of the most semantically discriminable feature set for a concept set, given a feature library.

To compute max capacity for a given concept set, we solve an assignment problem using the balanced merit function (Section 2.3) over the entire feature library. This yields a feature set. We define max capacity as the (generalized) semantic distance of this feature set for the given concept set. High max capacity indicates that the feature library contains at least one feature set with high semantic discriminability for the concept set. Low max capacity indicates no such feature set exists for that concept set, at least given the feature library.

In the case of two concepts, the balanced merit approach for computing max capacity gives the same result as exhaustively computing the semantic distance for each pair of colors, then finding the maximum of those semantic distances. Using balanced merit, though, allows max capacity to scale easily; it can be efficiently computed for large concept sets and feature sets. We also explored alternative ways to operationalize capacity (see Supplementary Material Section S.4).

3.4 The theory

Semantic discriminability theory posits that the capacity to produce semantically discriminable perceptual features for a set of concepts depends on the difference in feature-concept association distributions over a feature library. Briefly, distribution difference predicts capacity, distinct from the contribution of specificity. This idea differs from previous approaches, which primarily focused on color-concept associations for concepts in isolation when evaluating the potential to meaningfully encode particular concepts using color [20, 33].

Figure 1 shows the distinction between distribution difference and specificity of color-concept associations, with respect to capacity. It includes concept sets with large, medium, and small distribution differences. Capacity is illustrated with histograms below each concept set. They show the frequency of color sets across values of semantic distance (2485 possible 2-color sets from the UW-71 color library), with an arrow pointing at maximum semantic distance. Concept sets with large, medium, and small distribution differences result in high, medium, and low capacity, respectively. Yet, the concepts with medium capacity (driving and comfort) have far lower specificity than concepts with low capacity (eggplant and grape). The reason that concepts with low specificity can result in higher capacity than concepts with high specificity is that semantic discriminability depends on the difference in merit of each possible set of feature-concept assignments, not just isolated feature-concept associations (Section 2.3).

Figure 4 further illustrates this point with hypothetical color-concept association distributions for 2-concept sets that have higher capacity (top row) and lower capacity (bottom row). The colored dots on the distributions indicate the optimal assignment according to balanced merit (though this is arbitrary when the distributions are parallel because all assignments are equally poor). Next to each distribution pair is a histogram of semantic distances (as in Figure 1) and a bipartite graph for the colors with maximum semantic distance (thicker edges connecting colors and concepts indicate greater merit). Semantic distance (ΔS) is indicated below the bipartite graphs, and can be visually inspected by comparing the total merit of the outer edges vs. inner edges and assessing the degree to which one sum is larger. When distribution difference is high (top row), capacity is high, even if one concept has a uniform distribution (i.e., no specificity). However, when distribution difference is lower (bottom row), capacity is lower, even if both concepts have high specificity.

We chose the particular examples in Figure 1 and Figure 4 to highlight the dissociation between distribution difference and specificity, but we systematically tested for effects of these factors on capacity in Experiment 1.

4 Experiment 1

Experiment 1 tested the hypothesis that capacity for semantic discriminability is predicted by distribution difference, independent of specificity. We first collected color-concept association data from human participants, and used those data to calculate capacity, distribution difference, and specificity. We then tested our hypothesis on 2-concept sets (Section 4.2.1) and 4-concepts sets (Section 4.2.2). Semantic discriminability predicts people's ability to interpret palettes in visualizations [31], so our modeling approach for understanding capacity for semantic discriminability should have implications for interpretability. The code and data for all experiments is at: `https://github.com/SchlossVRL/sem_disc_theory`.

4.1 Methods

4.1.1 Participants

185 undergraduates participated for credit in Introductory Psychology (mean age =18.66, 99 females, 86 males, gender provided through free-response). All gave informed consent and the UW–Madison IRB approved the protocol. Color vision was assessed by asking participants if they had difficulty distinguishing between colors relative to the average person and if they considered themselves colorblind. Participants were excluded if they answered yes to either (5 excluded).

4.1.2 Design, Displays, and Procedure

Participants judged the association between each of 71 colors and each of 20 concepts. The colors were the UW-71 color library, an extension of the UW-58 colors [31], see Supplementary Material for details and Table S.1 for CIELAB coordinates.[3] The concepts were from Lin et al. [20], including 5 concepts in each of four concept categories (fruits, vegetables, activities, and properties) (Table 1). Participants were randomly assigned to one of four categories (fruits $n = 46$, vegetables $n = 45$, activities $n = 45$, properties $n = 44$). They judged all colors for all five concepts within their assigned category (71 colors $\times$ 5 concepts $= 355$ trials). Trials were presented in a blocked randomized design—all colors were presented in a random order for a given concept before starting the next concept, and concept order was also randomized.

The displays included the concept word centered at the top of the screen (font-size: 24 pt, font-family: Lato) and colored square centered below (80 px $\times$ 80 px). Below the colored square, was a line-mark slider scale (400 px long), with the left end labeled "not at all" and the right end labeled "very much" and the center marked with a vertical line (3 px wide and 32 px tall). The background was gray (CIE Illuminant D65, x = .3127, y = .3290, Y = 10 cd/m^2), so that very dark colors (e.g., black) and very light colors (e.g., white) could be seen against the background. Data were recorded in pixel units, and scaled to range from 0-1. Displays were generated using the jsPsych JavaScript library [9], presented on participants' personal devices.

Participants were told they would see a set of concepts and series of colors, one concept and color at a time. Their task was to rate how much they associated the color with the concept by moving the slider on the scale from "not at all" to "very much", and clicking "next" to continue. Before beginning, they were shown a list of all concepts and the UW-71 colors. They were asked to anchor the endpoints of the rating scale for each concept [26] by thinking about which color they associated the most/least with that concept, and considering these colors as representing the ends of the slider scale for that concept. During the experiment, ratings were blocked by concept, and after each block participants were told how many blocks remained.

4.2 Results and Discussion

4.2.1 2-Concept sets

We began by calculating the mean color-concept association ratings over participants. Next, for all $k = 2$ concepts out of the $n = 20$ concepts in Table 1 (190 2-concept sets in total), we used the mean color-concept associations to calculate capacity for semantic discriminability, distribution difference, and mean specificity. To calculate capacity, we followed the method in Section 3.3. To calculate distribution difference, we used total variation (TV) in Equation (4) and normalized the TV values to range from 0 to 1. To calculate mean specificity, we first computed entropy (H) for each concept (Equation (3)) over $N = 71$ colors, and then computed the mean entropy over concepts within each set. Given that higher specificity corresponds to lower entropy, we normalized mean entropy to range from 0 to 1 and subtracted the scores from 1, such that larger numbers indicated higher specificity. Figure S.2 in the Supplementary Material shows the raw entropy for each concept. Concepts with lower entropy/higher specificity corresponded to colorable concepts in [20], and concepts with higher entropy/lower specificity corresponded to non-colorable concepts in [20].

Table 1: Full set of concepts in Experiment 1 (first four columns of concepts were used in Experiment 2).

Category	Concepts				
Fruits	peach	cherry	grape	banana	apple
Vegetables	corn	carrot	eggplant	celery	mushroom
Activities	working	leisure	sleeping	driving	eating
Properties	efficiency	speed	safety	comfort	reliability

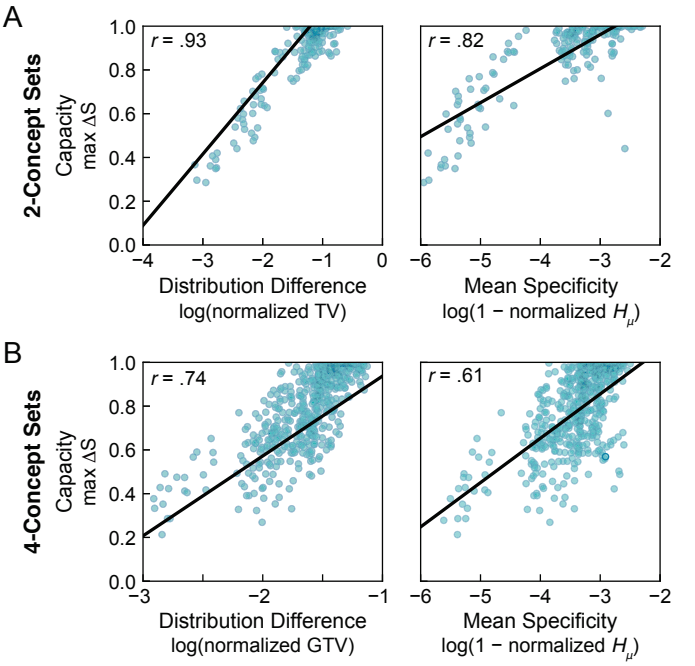

Figure 5: Relations between capacity for semantic discriminability and distribution difference (log(normalized (generalized) total variation distance); left) and specificity (log(1− normalized mean entropy); right) for 2-concept sets (top) and 4-concept sets (bottom). For 4-concept sets we downsampled from 4845 points to 500 points to avoid overplotting.

Figure 5A shows the relation between capacity for semantic discriminability and distribution difference (left), and mean specificity (right). For both distribution difference and mean specificity, we plotted the log of the normalized scores to preserve linearity. The correlation between capacity and distribution difference over all 190 2-concept sets was strongly positive ($r(188) = .93$, $p < .001$), with a strong trend for capacity to increase with increased distribution difference. The correlation between capacity and mean specificity was also significantly positive ($r(188) = .82$, $p < .001$), but was significantly weaker than the correlation with distribution difference (Fisher's r-to-z transformation $z(188) = 4.85$, $p < .001$). This weaker correlation can be attributed, in

[3] We converted CIELAB to RGB using MATLAB's lab2rgb function, which makes assumptions about monitor characteristics, so the colors were not exact renderings of CIELAB coordinates. Without calibration, the colors rendered by RGB coordintes may vary across monitors, but using a fixed correspondence between D65 CIELAB and RGB can approximate intended colors online [36].

Table 2: Multiple linear regression predicting capacity for semantic discriminability from distribution difference and mean specificity for all 2-concept sets and 4-concept sets.

Model	Factor	β	SE	t	p
2-concept	Intercept	.867	.005	181.9	$< .001$
	Distribution diff.	.160	.010	15.6	$< .001$
	Specificity	.002	.010	.201	.841
4-concept	Intercept	.772	.002	483.8	$< .001$
	Distribution diff.	.235	.004	53.6	$< .001$
	Specificity	$-.112$	.004	-25.5	$< .001$

part, to there being concept sets with high capacity, despite moderate to low mean specificity, and concept sets with low capacity despite high mean specificity (Figure 5, right).

To examine whether distribution difference and mean specificity contributed independently to capacity, we used a multiple linear regression model to predict capacity from these two factors (z-scored to center them and put them on the same scale). As shown in Table 2, distribution difference was a strong significant predictor, and mean specificity was not significant. Thus, the variance explained in capacity by distribution difference was independent from mean specificity, and mean specificity did not contribute after accounting for distribution difference.

4.2.2 4-Concept sets

For all $k = 4$ concepts out of the $n = 20$ concepts in Table 1 (4845 4-concept sets in total), we used the mean color-concept associations to calculate capacity, distribution difference, and mean specificity, as described in Section 4.2.1 for 2-concept sets. However, instead of semantic distance to compute capacity we used generalized semantic distance (Section 3.2), and instead of using TV to compute distribution difference, we used GTV (Equation 5, Section 3.1.2).

Figure 5B shows the relation between capacity for semantic discriminability and distribution difference (left), and mean specificity (right) for 4-concept sets. As for 2-concept sets, we used the log of the normalized distribution difference and mean specificity scores to preserve linearity. Capacity was positively correlated with both distribution difference ($r(4843) = .74$, $p < .001$) and mean specificity ($r(4843) = .61$, $p < .001$), but the correlation with distribution difference was greater (Fisher's r-to-z transformation ($z(4843) = 11.88$, $p < .001$).

Using the same regression analysis as for 2-concept sets, distribution difference was a strong significant predictor (Table 2). Mean specificity a weak significant predictor, but surprisingly it was negative, such that less specificity resulted in greater capacity in the context of this model.

In summary, Experiment 1 supports the hypothesis that the capacity to produce semantically discriminable color palettes for a set of concepts depends on the difference in color-concept association distributions, independent of specificity. Considering specificity of color-concept associations in isolation is insufficient. These results emphasize the importance of considering relative color-concept associations for a given set of concepts, rather than the concepts in isolation, when evaluating the potential for semantically discriminable color palettes.

5 Experiment 2

Semantic discriminability theory implies that if concept sets have high capacity for semantic discriminability, it should be possible to create encoding systems assigning those concepts to colors that people can interpret. People should be able to interpret the correct mappings between colors and concepts, even for concepts previously considered "non-colorable," insofar as the colors are semantically discriminable. We tested this hypothesis in Experiment 2. We defined accuracy as the proportion of responses that matched the optimal mapping specified by an assignment problem using the balanced merit function (see Section S.7 in the Supplementary Material for a further discussion on accuracy, and its relation to measures of semantic discriminability).

5.1 Methods

5.1.1 Participants

98 participants (74 males, 24 females) were recruited on Amazon Mechanical Turk. All gave informed consent, and the UW–Madison IRB approved the protocol. Eight were excluded for not reaching 100% accuracy on catch trials (Section 5.1.2), three of which reported atypical color vision. All other participants reported typical color vision.

5.1.2 Design, Displays, and Procedure

For each trial, participants were presented with a bar graph centered on the screen, consisting of four colored bars (Figure 6A). Each bar was 130 px wide and varied in height randomly (from 260-300 px high). The bars were spaced 45 px apart. At the start of the trial, a set of four concepts (22 pt font) was centered above the graph in a random order. The y-axis was unlabeled. Below the x-axis, there were empty boxes 120-px wide and 50 px high. During the trial, participants labeled each bar by clicking on a label and dragging/dropping it in the empty box below the bar. The displays were generated using the Charts.js and jsPsych JavaScript libraries.

Each participant completed 64 trials, which included 8 color-concept sets $\times$ 8 color positionings within each set. Figure 6B shows the palettes for each set. The stimuli were constructed using displays like in Figure 6A, but swapping out the concept sets and corresponding color palettes, and balancing the bar color positioning as follows.

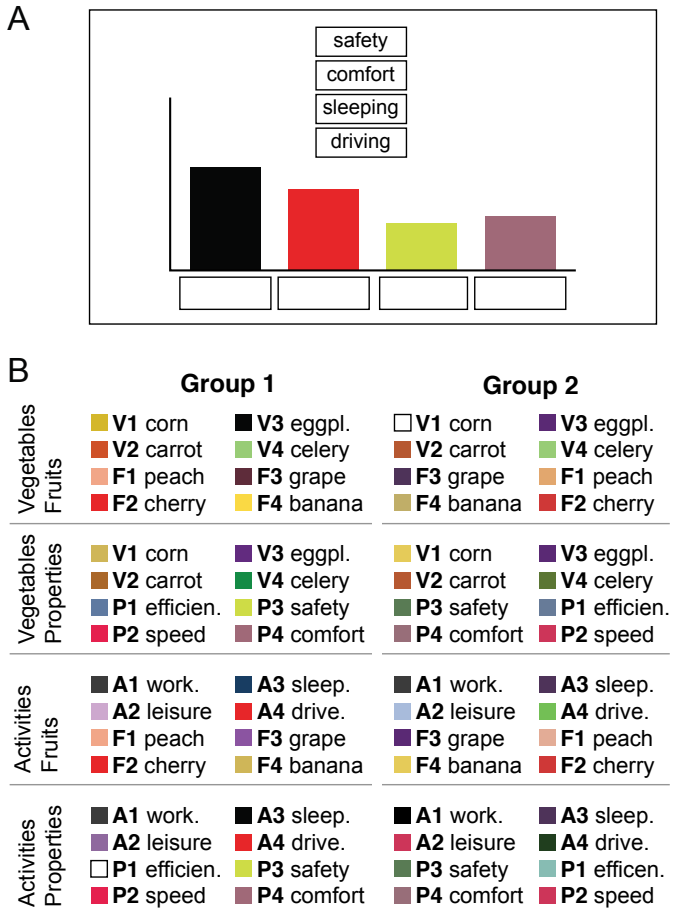

Figure 6: (A) Example trial in Experiment 2. Participants labeled each bar by clicking the label and dragging/dropping it in the box below the bar. (B) Palettes and corresponding concepts used to construct the stimuli (see text for details).

Concept sets. To generate the concept sets, we randomly selected four concepts from each of the concept categories from Experiment 1 (fruits (F), vegetables (V), activities (A), properties (P)) and labeled

them 1-4 (Table 1, Figure 6B). We then tied pairs of concepts within each category (e.g., V1-V2, V3-V4). We combined pairs of concepts such that all participants saw (1) vegetables with fruits, (2) vegetables with properties, (3) activities with fruits, and (4) activities with properties. Using this design, we created two groups of stimuli, divided over two groups of participants to reduce the number of trials for any one participant. Group 1 saw sets of four concepts, with concepts 1 and 2 in one category paired with concepts 1 and 2 in the other category (e.g., V1-V2 with F1-F2), and sets of four concepts with concept 3 and 4 in one category paired with concepts 3 and 4 in the other category (e.g., V3-V4 with F3-F4). Group 2 saw the opposite pairings (e.g., V1-V2 with F3-F4, V3-V4 with F1-F2). Within this design, all participants saw each concept an equal number of times. Participants were randomly assigned to Group 1 ($n = 47$) or Group 2 ($n = 43$).

Color palettes. For each concept set, we generated its color palette using the balanced merit function (Equation (1)) in an assignment problem. The resulting assignments determined the encoded mapping we defined as "correct." We used the balance merit function because previous evidence suggested it aligns with the merit people use in assignment inference (see Section 2.3). Merit was computed over the color-concept association data reported in Experiment 1 for all 71 colors in the UW-71 library [4] The color palettes are shown in Figure 6B. The CIELAB coordinates for the palette colors can be found at `https://github.com/SchlossVRL/sem_disc_theory`. The graphs were presented on a gray background approximating CIE Illuminant D65 ($x = .3127$, $y = .3290$, $Y = 10$, cd/m^2).

Bar color positioning. Each of the eight color-concept sets for a given group (Figure 6B) was presented eight times in eight bar color positionings along the x-axis. This was done using a blocked randomized design, so all eight color-concept sets appeared once in a random order, randomly assigned to a color positioning within a block, before starting the next block. The eight possible color positionings were defined using a Latin square design (four positionings, left/right reversed). Thus, within a color set, each color appeared in each of four positions twice, with the colors to its left/right in opposite positions.

Catch trials. We included eight catch trials, one per block, in which bars were colored a shade of red, yellow, green, and blue, and the labels were "red", "yellow", "green", and "blue." We set an *a priori* exclusion criterion that participants must be 100% accurate on these catch trials, otherwise their data would be excluded from analysis.

Participants were told they would see a series of colored bar graphs, with four bars and four words at the top of the screen. Their task was to match each word to its corresponding bar color by clicking and dragging the label to the empty box below the bar. They were told to use their best guess if they were unsure how to match the labels to the bar colors. They then completed a practice trial with four concepts that were not in the main experiment (blueberry, mango, strawberry, lemon) and colors chosen by the balanced merit function. Associations for these concepts had been collected for a different project. During the trials, all bars had to be labeled before a "continue" button could be pressed to go to the next trial. Once placed in a box, a label could be dragged to another box and all labels could be reset to the starting position by pressing a "reset label" button. Trials were separated by a 100 ms. inter-trial interval. Participants received breaks after each block, and were told the proportion of completed trials at each break.

5.2 Results and Discussion

For each participant, we calculated the proportion of times they chose each concept for each color in each color-concept set, averaged over bar color positioning. These results are shown for a subset of the concept sets in Figure 7A (top row), and for all concepts sets in Figure S.6. For each color-concept pairing, we calculated accuracy as the proportion of trials in which participants selected the optimal pairing (defined with respect to balanced merit) (Section S.7). The arrows below the x-axis in Figures 7A and S.6 point up to the correct color.

[4]Due to a scaling issue during palette creation, 15 of the 64 color-concept pairings were not optimal. This did not affect the analyses, but accuracy may have been greater if participants had seen fully optimized palettes.

Table 3: Logistic mixed-effect model predicting accuracy from specificity of the concept, semantic contrast of the concept's correct color, and association between the concept and its correct color.

Fixed Effects	β	**SE**	z	p
Intercept	1.272	.176	7.249	$< .001$
Specificity	.226	.088	2.577	.001
Semantic Contrast	.645	.081	7.926	$< .001$
Association Strength	$-.064$	.057	-1.12	.262

We first tested whether concept sets with higher capacity enabled creating encoding systems that were easier to interpret. To do so, we correlated max capacity for each of the 16 concept sets with mean accuracy over all colors within each set. There was a significant relation ($r = .58, p < .02$), indicating greater capacity for semantic discriminability corresponded to greater interpretability.

Next, we tested whether participants' patterns of color choices for each concept were correlated with model predictions computed by solving an assignment problem with perturbed association ratings (the Monte Carlo process described in Section 3.2 over 1000 iterations). These predictions are shown in the bottom row of Figure 7A and in Figure S.6. In the model predictions, the height of the bars correspond to the proportion of times each color was assigned to each concept. The predictions strongly correlated with participant responses over the full dataset of 4 colors $\times$ 16 4-concept sets ($r(126) = .95, p < .001$), with high correlations for each group (Group 1: $r(126) = .96$, Group 2: $r(126) = .94, ps < .001$).

Finally, we tested our hypothesis that participants would be able to interpret the correct mappings between individual colors and concepts, insofar as the colors were semantically discriminable. Figure 7A shows that participants chose the correct colors well above chance, even for concept sets in which all concepts have been called non-colorable (e.g., {sleeping, driving, safety, speed}. To examine whether accuracy for given a concept varied depending on semantic discriminability of its correct color, in 7B, we plotted accuracy for each concept as a function of the semantic contrast of its correct color (see Section 3.2 and Section S.7 for details on semantic contrast). Plots are separated by concept category, with four points per concept, corresponding to the four color-concept sets in which it appeared. Generally, the slopes of the best fit lines for each concept were positive, indicating that accuracy increased with semantic contrast. Responses for some concepts (e.g., fruits) were highly accurate for all color-concept sets because their optimal colors have high semantic contrast in all concept sets we tested.

We analyzed this pattern of accuracies using a mixed-effect logistic regression model predicting accuracy for each concept in each set using three factors: semantic contrast of the correct color for that concept (relative to the other colors in the palette), specificity of the concept as defined in Experiment 1, and association strength between the concept and its correct color (previously shown to influence accuracy in similar tasks [31, 32]). These predictors were calculated using data from Experiment 1 (different participants from Experiment 2). We also included by-subject random intercepts and by-subject random slopes for each factor. We z-scored the individual predictors to put them on the same scale and set the correlations between the random slopes to be 0 to help the model converge. As shown in Table 3, accuracy significantly increased with greater semantic contrast and with greater specificity. Association strength was not significant.

Overall, accuracy was greater for concepts previously considered colorable (fruits and vegetables) ($M = 0.76$, $SD = 0.23$) than those considered non-colorable (activities and properties) ($M = 0.56$, $SD = 0.24$) (Figure 7B). But, all activities and properties had at least one instance that was as accurate as fruits and vegetables, and all instances were above chance. Moreover, accuracy for a given concept varied based on semantic contrast with its correct color, which cannot be explained by specificity of the concept in isolation. These results suggest that any concept has potential to be meaningfully encoded using color if the color has sufficient semantic contrast with other colors in the palette.

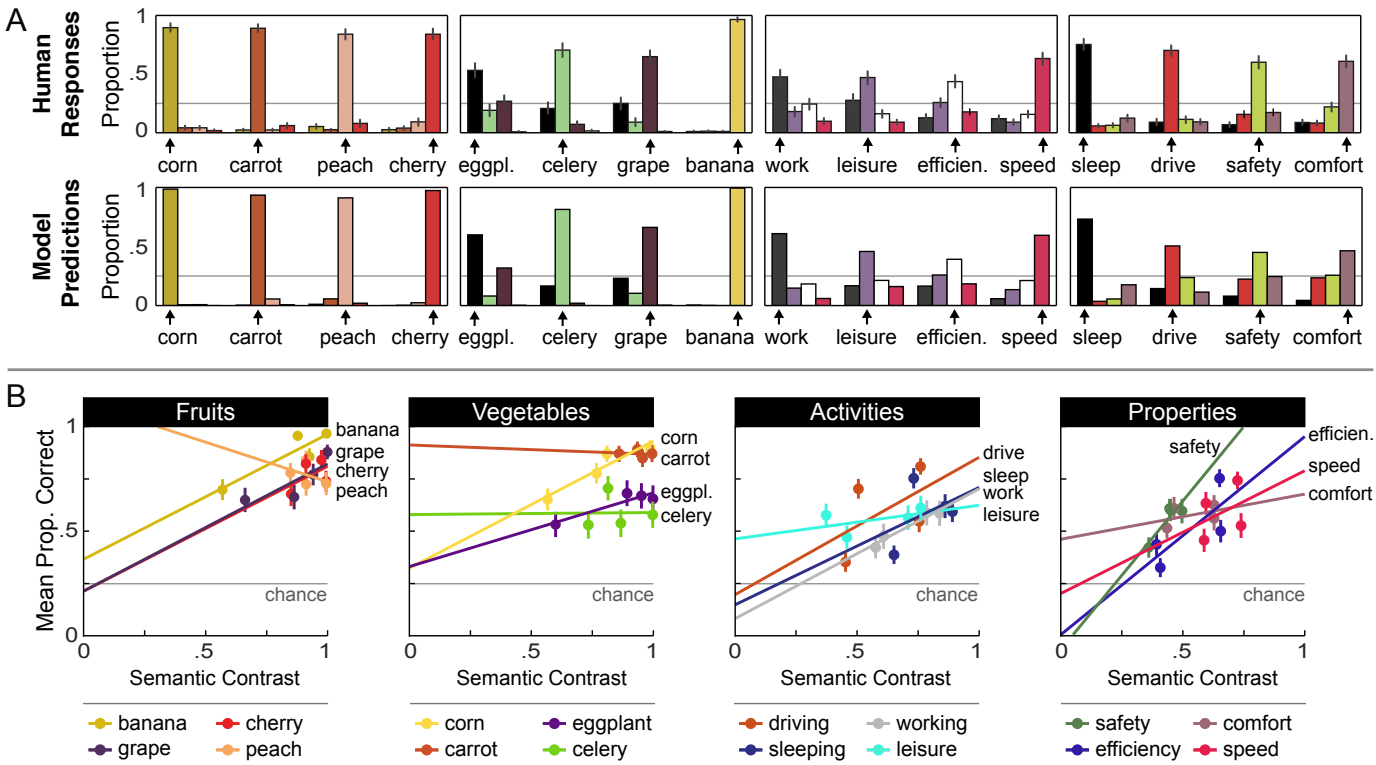

Figure 7: (A) Proportion of times participants chose each color (top) and predicted proportions from generalized semantic distance (bottom) for a subset of palettes from group 1 (see Figure S.6 in the Supplementary Material for the full dataset). The correct response for each concept is marked by an arrow along the x-axis. The colors of the bars correspond to the colors of the stimuli. (B) Mean proportion correct for each concept in each palette as a function of semantic contrast of its correct color (best fit lines drawn for each concept). All the points for a given concept and corresponding best fit line are shown in the same color to help group the points in this figure (these colors were not necessarily the colors shown in the experiment. In (A) and (B) gray horizontal lines correspond to chance (.25) and error bars represent ± standard errors of the means.

6 General Discussion and Conclusion

In this paper we presented semantic discriminability theory to specify constraints on producing semantically discriminable perceptual features for visual communication. The theory states that capacity for creating semantically discriminable features for a concept set is constrained by the difference in feature-concept association distributions for those concepts. Supporting the theory, Experiment 1 showed that distribution difference between color-concept association distributions predicted capacity for semantic discriminability in 2- and 4-concept sets, independent of specificity. And, Experiment 2 indicated people can correctly interpret mappings for concepts previously considered non-colorable, but their ability to do so depended on semantic contrast with respect to the other colors in the encoding system.

Semantic discriminability theory is rooted in feature-concept associations, which can vary cross-culturally [16, 17, 38]. The theory implies that distribution difference will predict capacity for semantic discriminability in different cultures, as long as the association distribution data reflect the associations held by a given culture.

The theory further implies that any factor that influences distribution difference for a set of concepts can affect capacity. Below, we propose criteria for producing distribution differences that support adequate capacity for semantic discriminability. Evaluating these criteria will help guide future work on the potential and limitations of semantic discriminability for colors and for other perceptual features.

Criterion 1: Need for *some* specificity. At least *some* concepts in the concept set must have association distributions with some specificity. If all concepts in a set have uniform distributions, there will be no capacity for semantic discriminability (Figure 4). Some perceptual features may not support specificity as well as color does, such as line orientation. If so, such features might be less useful for communicating meaning in information visualizations.

Criterion 2: Need for feature library variability. To be sensitive to differences in feature-concept associations, if they exist, the feature library must be sufficiently variable. In color, variability is achieved by sampling widely over color space, as opposed to sampling say, only the bluish part of the space. One can systematically sample over color spaces because color spaces are well-defined feature sources. But, such sampling may pose a challenge for less well-specified feature sources (e.g., all possible shapes or all possible textures).

Criterion 3: Need for large enough feature library. The feature library must be large enough to detect small, but important differences between feature-concept association distributions. E.g., a library with only two colors, a blue and red, might be large enough to produce distinct association distributions for the concepts sky and rose, but a library with more colors (e.g., more shades of blue) would be needed to produce distinct distributions for concepts like noon sky and night sky.

Conclusion. We presented and evaluated semantic discriminability theory to define constraints on creating semantically discriminable features for perceptual encoding systems. The theory implies that any concept has potential to be meaningfully encoded using color, if the criteria above are met. Thus a concept that has low specificity (i.e., uniform distribution), can meaningfully be encoded by a color, if other concepts in the set have sufficiently different distributions. This is possible because people infer globally optimal mappings between colors and concepts, even if that means inferring concepts map to weakly-associated colors. The theory implies, and our results suggest, color is more robust for visual communication than previously thought.

Acknowledgments

We thank Rob Nowak, Melissa Schoenlein, Kevin Lande, Tim Rogers, Chris Thorstenson, Anna Bartel, and Maureen Stone for helpful discussions. This project was supported by the UW–Madison Office of the Vice Chancellor for Research and Graduate Education, Wisconsin Alumni Research Foundation, and NSF (BCS-1945303 to KBS).

REFERENCES

[1] C. L. Anderson and A. C. Robinson. Affective congruence in visualization design: Influences on reading categorical maps. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[2] L. Bartram, A. Patra, and M. Stone. Affective color in visualization. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1364–1374. ACM, 2017.

[3] B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, 1969.

[4] J. Bertin. *Semiology of graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, Madison, 1983.

[5] R. J. Brockmann. The unbearable distraction of color. *IEEE Transactions on Professional Communication*, 34(3):153–159, 1991.

[6] R. Burkard, M. Dell'Amico, and S. Martello. *Assignment Problems: revised reprint*. SIAM, 2012.

[7] M. A. Changizi, Q. Zhang, and S. Shimojo. Bare skin, blood and the evolution of primate colour vision. *Biology Letters*, 2(2):217–221, 2006.

[8] B. R. Conway. Color vision, cones, and color-coding in the cortex. *The Neuroscientist*, 15(3):274–290, 2009.

[9] J. R. De Leeuw. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1):1–12, 2015.

[10] A. J. Elliot, M. A. Maier, A. C. Moller, R. Friedman, and J. Meinhardt. Color and psychological functioning: the effect of red on performance attainment. *Journal of Experimental Psychology: General*, 136(1):154, 2007.

[11] H. Fang, S. Walton, E. Delahaye, J. Harris, D. Storchak, and M. Chen. Categorical colormap optimization with visualization case studies. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):871–880, 2017.

[12] M. Hasantash, R. Lafer-Sousa, A. Afraz, and B. R. Conway. Paradoxical impact of memory on color appearance of faces. *Nature Communications*, 10:1–10, 2019.

[13] C. Havasi, R. Speer, and J. Holmgren. Automated color selection using semantic knowledge. In *2010 AAAI Fall Symposium Series*, 2010.

[14] N. Humphrey. The colour currency of nature. In T. Porter and B. Mikellides, editors, *Colour for Architecture Today*, pages 95–98. Taylor & Francis, 1976.

[15] A. Jahanian, S. Keshvari, S. Vishwanathan, and J. P. Allebach. Colors–messengers of concepts: Visual design mining for learning color semantics. *ACM Transactions on Computer-Human Interaction*, 24(1):2, 2017.

[16] D. Jonauskaite, A. M. Abdel-Khalek, A. Abu-Akel, A. S. Al-Rasheed, J.-P. Antonietti, Á. G. Ásgeirsson, K. A. Atitsogbe, M. Barma, D. Barratt, V. Bogushevskaya, et al. The sun is no fun without rain: Physical environments affect how we feel about yellow across 55 countries. *Journal of Environmental Psychology*, 66:101350, 2019.

[17] D. Jonauskaite, J. Wicker, C. Mohr, N. Dael, J. Havelka, M. Papadatou-Pastou, M. Zhang, and D. Oberfeld. A machine learning approach to quantify the specificity of colour–emotion associations and their cultural differences. *Royal Society Open Science*, 6(9):190741, 2019.

[18] P. Kay, N. J. Smelser, P. B. Baltes, and B. Comrie. *The Linguistics of Color Terms*. Citeseer, 2001.

[19] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[20] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer. Selecting semantically-resonant colors for data visualization. In *Computer Graphics Forum*, volume 32, pages 401–410. Eurographics Conference on Visualization, 2013.

[21] A. Lindner, N. Bonnier, and S. Süsstrunk. What is the color of chocolate?–extracting color values of semantic expressions. In *Conference on Colour in Graphics, Imaging, and Vision*, volume 2012, pages 355–361. Society for Imaging Science and Technology, 2012.

[22] A. Lindner, B. Z. Li, N. Bonnier, and S. Süsstrunk. A large-scale multilingual color thesaurus. In *Color and Imaging Conference*, volume 2012, pages 30–35. Society for Imaging Science and Technology, 2012.

[23] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

[24] F. Nielsen and R. Bhatia. *Matrix Information Geometry*. Springer Berlin Heidelberg, 2012.

[25] L.-C. Ou, M. R. Luo, A. Woodcock, and A. Wright. A study of colour emotion and colour preference. Part I: Colour emotions for single colours. *Color Research & Application*, 29(3):232–240, 2004.

[26] S. E. Palmer, K. B. Schloss, and J. Sammartino. Visual aesthetics and human preference. *Annual Review of Psychology*, 64:77–107, 2013.

[27] R. Rathore, Z. Leggon, L. Lessard, and K. B. Schloss. Estimating color-concept associations from image statistics. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1226–1235, 2020.

[28] A. H. Robinson. *The Look of Maps*. University of Wisconsin Press, Madison, 1952.

[29] K. B. Schloss. A color inference framework. In . G. V. P. L. MacDonald, C. P. Biggam, editor, *Progress in Colour Studies: Cognition, Language, and Beyond*. John Benjamins, Amsterdam, 2018.

[30] K. B. Schloss, C. C. Gramazio, A. T. Silverman, M. L. Parker, and A. S. Wang. Mapping color to meaning in colormap data visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):810–819, 2019.

[31] K. B. Schloss, Z. Leggon, and L. Lessard. Semantic discriminability for visual communication. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1022–1031, 2021.

[32] K. B. Schloss, L. Lessard, C. S. Walmsley, and K. Foley. Color inference in visual communication: the meaning of colors in recycling. *Cognitive Research: Principles and Implications*, 3(1):5, 2018.

[33] V. Setlur and M. C. Stone. A linguistic approach to categorical color assignment for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):698–707, 2016.

[34] P. Shah and J. Hoeffner. Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, 14(1):47–69, 2002.

[35] S. C. Sibrel, R. Rathore, L. Lessard, and K. B. Schloss. The relation between color and spatial structure for interpreting colormap data visualizations. *Journal of Vision*, 20(12):7–7, 2020.

[36] M. Stone, D. A. Szafir, and V. Setlur. An engineering model for color difference as a function of size. In *Color and Imaging Conference*, volume 2014, pages 253–258. Society for Imaging Science and Technology, 2014.

[37] J. W. Tanaka and L. M. Presnell. Color diagnosticity in object recognition. *Perception & Psychophysics*, 61(6):1140–1153, 1999.

[38] D. S. Y. Tham, P. T. Sowden, A. Grandison, A. Franklin, A. K. W. Lee, M. Ng, J. Park, W. Pang, and J. Zhao. A systematic investigation of conceptual color associations. *Journal of Experimental Psychology: General*, 149(7):1311–1332, 2020.

[39] C. A. Thorstenson, A. J. Elliot, A. D. Pazda, D. I. Perrett, and D. Xiao. Emotion-color associations in the context of the face. *Emotion*, 18(7):1032–1042, 2018.

[40] C. Witzel, H. Valkova, T. Hansen, and K. R. Gegenfurtner. Object knowledge modulates colour appearance. *i-Perception*, 2(1):13–49, 2011.

[41] G. Wyszecki and W. S. Stiles. *Color Science*. Wiley New York, 1982.