Semantic Discriminability for Visual Communication

Karen B. Schloss, Zachary Leggon, and Laurent Lessard

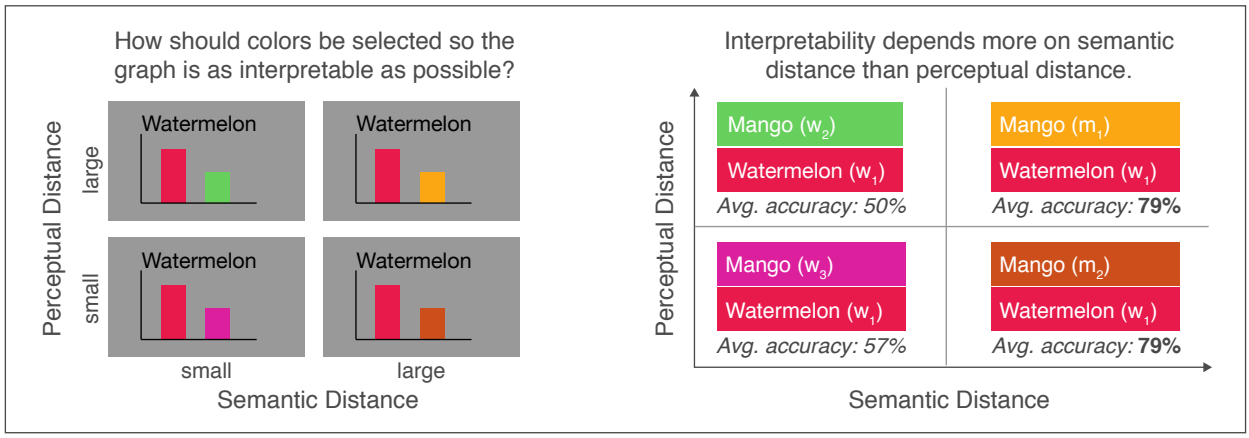

Figure 1. This study investigated how varying perceptual and semantic distance influenced people's ability to interpret color palettes. Left: four example trials from Experiment 2, in which participants reported which bar (left/right) corresponded to the target fruit named above the graph (either mango or watermelon). No legend was provided, so participants had to rely on their inferred mappings to complete the task. Right: aggregated results corresponding to the color pairs from the left panel. The colors are labeled with their correct assignments (see Section 3.2.1 for details) as well as the color names in parentheses (see Table S.1 for coordinates). Participants had higher mean accuracy when the *semantic distance* between the pair of colors was greater (a notion we define in Section 3.2.2), whereas perceptual discriminability had little effect on accuracy when controlling for semantic distance.

Abstract—To interpret information visualizations, observers must determine how visual features map onto concepts. First and foremost, this ability depends on perceptual discriminability; observers must be able to see the difference between different colors for those colors to communicate different meanings. However, the ability to interpret visualizations also depends on semantic discriminability, the degree to which observers can infer a unique mapping between visual features and concepts, based on the visual features and concepts alone (i.e., without help from verbal cues such as legends or labels). Previous evidence suggested that observers were better at interpreting encoding systems that maximized semantic discriminability (maximizing association strength between assigned colors and concepts while minimizing association strength between unassigned colors and concepts), compared to a system that only maximized color-concept association strength. However, increasing semantic discriminability also resulted in increased perceptual distance, so it is unclear which factor was responsible for improved performance. In the present study, we conducted two experiments that tested for independent effects of semantic distance and perceptual distance on semantic discriminability of bar graph data visualizations. Perceptual distance was large enough to ensure colors were more than just noticeably different. We found that increasing semantic distance improved performance, independent of variation in perceptual distance, and when these two factors were uncorrelated, responses were dominated by semantic distance. These results have implications for navigating trade-offs in color palette design optimization for visual communication.

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

1 Introduction

In visual communication, designers produce information visualizations by encoding concepts in visual features, and observers interpret visualizations by decoding concepts from visual features [4, 39]. Interpreting visualizations involves multiple component processes, including (1) perceiving and identifying important features within a visualization, (2) mapping those features to the concepts they represent, and (3) deriving implications about information represented in the visualization [32]. For example, to interpret visualizations that encode categories using color (e.g., bar graphs, choropleth maps, transit maps, and recycling bin signage), observers must perceive and distinguish the colors, determine how each color in the palette maps to a category, and then use that mapping to glean knowledge from the visualization (e.g., patterns of data from bar graphs or choropleth maps, which train to take from a transit map, or where to discard paper from recycling signage).

Many factors influence interpretability, including (1) characteristics of visualizations, (2) observers' knowledge about visualizations, and (3) observers' knowledge about content in visualizations (outlined in [32], with respect to graphs). Here, we focus on how interpretability is influenced by perceptual characteristics of visualizations that are used to encode meaning (e.g., colors) and observers' semantic associations with those perceptual features. Thus, for present purposes, we operationalize "interpretability" as the ability to accurately decode the encoded mapping. We aim to develop a deeper understanding of how interpretability is influenced by two properties: *perceptual discriminability* and *semantic discriminability*.

- *Karen B. Schloss, Psychology and Wisconsin Institute for Discovery, University of Wisconsin–Madison. Email: kschloss@wisc.edu.*
- *Zachary Leggon, Biology and Wisconsin Institute for Discovery, University of Wisconsin–Madison, Email: zleggon@wisc.edu.*
- *Laurent Lessard, Mechanical and Industrial Engineering, Northeastern University. Email: l.lessard@northeastern.edu.*

Manuscript received 30 Apr. 2020; revised 31 July 2020; accepted 14 Aug. 2020. Date of publication 26 Oct. 2020; date of current version 15 Jan. 2021. Digital Object Identifier no. 10.1109/TVCG.2020.3030434

1.1 Perceptual discriminability

Perceptual discriminability is the degree to which observers can perceive differences between different visual features (e.g., colors, sizes, shapes, or textures). Some amount of perceptual discriminability is necessary because observers cannot decode different meanings from perceptually identical features [17] (e.g., they must be able to perceive the difference between two shades of blue to decode that those blues encode different concepts). Thus, visualization research has emphasized the importance of understanding perceptual discriminability [8, 14, 33, 34], including how it varies with mark size [33] and shape [34]. And, design guidelines have emphasized the importance of representing categorical information with colors that are well-separated in color space [12, 14, 32]. If perceptually discriminable features are accompanied by verbal descriptions specifying the encoded mapping (e.g., legends or labels), then observers have all the information required to decode the encoded mapping. This rationale supports using pre-made color palettes (e.g., Tableau and Colorbrewer palettes [12]) that have been designed to ensure perceptual discriminability. However, the ability to decode encoded mappings depends on more than perceptual discriminability and legend reading, as explained below.

1.2 Semantic discriminability

We define semantic discriminability as the degree to which observers can infer a unique mapping between visual features and concepts, based on the visual features and concepts alone (i.e., without legends or labels). For example, if observers are given an unlabeled graph containing yellow and blue colored bars and are told the graph is about the concepts banana and blueberry, they could easily infer that yellow maps to banana and blue maps to blueberry. This is because yellow and blue are semantically discriminable for the concepts banana and blueberry. Conversely, it would be more difficult to infer how orangish-yellow and greenish-yellow map to the concepts banana and lemon because both colors are similarly associated with both concepts, and thus less semantically discriminable.

Semantic discriminability might sound similar to interpretability, but they are distinct constructs. Semantic discriminability concerns the ability to infer a unique mapping (irrespective of the encoded mapping), whereas interpretability concerns the ability to decode the correct mapping (specified by the encoded mapping). Building on the banana/blueberry graph example, yellow and blue would be semantically discriminable colors, regardless of the encoded mapping in the graph. Observers would infer that yellow maps to banana and blueberry maps to blue. Now, if the encoded mapping was yellow-banana/blue-blueberry, the graph would be easy to interpret because the encoded mapping matched the inferred mapping. But, if the encoded mapping was blue-banana/yellow-blueberry (i.e., cross-mapped [7]), the graph would be harder to interpret because the encoded mapping did not match the inferred mapping (i.e., Kosslyn's compatibility principle [17], Tversky et al.'s congruence principle [37]). Observers are better at interpreting colors in visualizations when encoded mappings match inferred mappings, even when there is a clear legend [19, 29].

Based on the examples above, one might conclude that interpretability depends only on association strengths of encoded color-concept pairs. Lin et al. [19] referred to palettes in which colors evoke the concepts they represent as *semantically resonant* color palettes. However, interpretability can also be achieved when not all color-concept pairs are semantically resonant [30], see Section 2.2. Rathore et al. [27] referred to this more general case as *semantically interpretable* color palettes. For simplicity, we use the term *interpretability* in the present work to refer to the more general case.

1.3 Perceptual vs. semantic discriminability?

From prior work, it is clear that interpretability hinges on some degree of perceptual discriminability [3, 14, 33, 34] and interpretability benefits from semantic discriminability [30]. However, given previous research, it is currently unknown whether increasing semantic discriminability improves interpretability, beyond that which can be explained by perceptual discriminability. Returning to our banana/blueberry/lemon examples used so far in this introduction, these examples were intended to build the intuition for semantic discriminability, but they confounded perceptual and semantic discriminability. Yellow and blue are both high in perceptual and semantic discriminability when encoding for the concepts banana and blueberry, and orangish-yellow vs. greenish-yellow are both low in perceptual and semantic discriminability when encoding for the concepts banana and lemon (assuming trichromatic color vision). Similarly, in prior work that suggested semantic discriminability improved interpretability, colors in the more semantically discriminable color palette were closer together in color space [30], see Section 2.2. So, it is unclear if this improvement was due to semantic or perceptual discriminability. Yet, semantic and perceptual discriminability *can* vary independently (Fig. 1), and understanding their independent effects on interpretability is important for determining how to resolve conflicts between them when optimizing color palette design.

Likewise, it is also unknown whether increasing perceptual discriminability beyond that which is needed for semantic discriminability influences interpretability. For two colors to be semantically discriminable, they must be sufficiently perceptually discriminable to tell them apart. Otherwise, observers could not reliably infer that one color maps more than another color does to a given concept. Thus, perceptual discriminability might not capture additional variance in interpretability beyond that which is explained by semantic discriminability.

To address these questions, we tested for independent effects of perceptual and semantic discriminability on interpretability. The results will not only provide a deeper understanding about the relative contributions of perceptual and cognitive factors for visual reasoning, but will also inform optimal color palette design. Designing effective palettes for information visualization requires navigating trade-offs between several, sometimes competing, factors (e.g., perceptual discriminability, semantic discriminability, name difference, emotional connotation, and aesthetics) [2, 8, 15, 19, 33]. Understanding the relative contribution of semantic and perceptual discriminability for interpretability will inform how to prioritize these factors when conflicts arise.

Contributions. Our study makes the following contributions. First, we define a metric called *semantic distance* for operationalizing semantic discriminability. Semantic distance depends on the relative association strengths between each color and each concept in the context of an encoding system (see Section 3.2). This is unlike perceptual distance, which only depends on the appearance of the two colors. The semantic distance between a given pair of colors may be large in the context of some concepts, but small in the context of other concepts.

Second, we present the results of two experiments that assess how perceptual distance and semantic distance influence interpretability. Evidence indicates that both factors can contribute to interpretability, but semantic distance dominates when the factors conflict. The results imply that increasing perceptual distance beyond that which is needed for semantic discriminability can improve interpretability, but when in conflict, priority should be given to maximizing semantic distance.

2 Background

When people interpret information visualizations, they do not simply absorb the displayed information in a bottom-up fashion. Instead, they have biases, or expectations, about how visual features map to meanings, which guide their interpretations. These biases span topics across the field of information visualization, including graphical perception [40, 41], visualizing uncertainty [28], and color [6, 19, 22, 29, 30]. Understanding and designing visualizations that align with these biases will help make visualizations that are easier to interpret [16, 24, 36]. In cases where this alignment may not be possible (i.e., multiple conflicting biases relevant to a particular visualization), an understanding of when expectations are violated can guide compensatory design decisions (e.g., extra labeling or verbal descriptions of the visualization).

Here, we focus on understanding expectations about assignments between colors and concepts for interpreting visualizations about categorical information. However, this discussion should apply to assignments between other perceptual features and concepts, as long as people have systematic associations between those features and concepts. In this section, we first describe how designers use *assignment problems* to produce encoded mappings between visual features and concepts. We

then present evidence that observers use assignment inference to decode encoded mappings when interpreting visualizations.

2.1 Assignment problems for encoding

Assignment problems have been used to create color palettes that optimize encodings between visual features and concepts [19, 30]. An assignment problem is a model for assigning items in one category (e.g., colors) to items in another category (e.g., concepts) in a manner that maximizes a total merit score [18, 23]. Assignment problems can be represented as bipartite graphs, as shown in Fig. 2. The square nodes are colors and the circular nodes are concepts. Edges are drawn between each color and each concept. The number on each edge represents the "merit score" of assigning that particular color to that concept, represented as $x_1, \ldots, x_4$. Merit scores can be calculated using different methods [19, 30] but the goal is always the same: construct a 1-to-1 assignment between each color and a concept, such that the sum of the merit scores of assigned color-concept pairs is maximized.

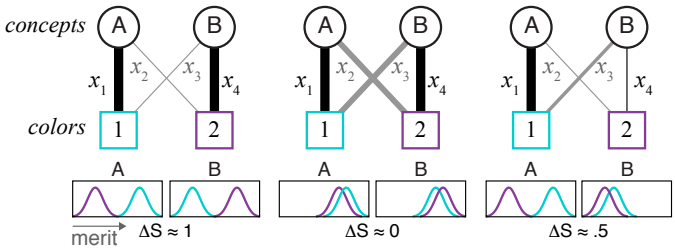

Figure 2. Bipartite graphs representing possible assignment problems between concepts (circles) and colors (squares). The *merit* of each pairing is represented by the thickness of edges connecting each color to each concept. Here, the merit is color-concept association strength. Plots below each bipartite graph show the relative merit distribution of each color for each concept, assuming that association strength is a normally distributed random variable. In all three examples, mean association strengths satisfy $x_1 + x_4 > x_2 + x_3$, so the outcomes of the assignment problems are the same: concept A is assigned with color 1 and concept B is assigned with color 2. Black lines indicate the chosen assignment and gray lines indicate the non-chosen assignment. ΔS indicates approximate semantic distance for the examples.

Assignment problems are deterministic. When there are two concepts and two colors as in Fig. 2, there are only two possible outcomes and the outcome is completely determined by the merit scores on each of the four edges. If the sum of the outer edges is greater than the sum of the inner edges ($x_1 + x_4 > x_2 + x_3$), then concept A is assigned to color 1 and concept B is assigned to color 2. If the sum of the inner edges is greater ($x_2 + x_3 > x_1 + x_4$), then concept A is assigned to color 2 and concept B to color 1. Fig. 2 illustrates bipartite graphs with three different patterns of merit on the edges, but all three scenarios produce identical outcomes: concept A is assigned to color 1 and concept B is assigned to color 2 because the merit scores satisfy $x_1 + x_4 > x_2 + x_3$. In the left and center bipartite graphs, the solution to the assignment problem ("global solution") matches the solution for each concept in isolation ("local solution", concept A is more associated with color 1 than color 2, and concept B is more associated with color 2 than color 1). However, in the rightmost bipartite graph, the local and global solutions conflict: concept B is more associated with color 1, yet it is assigned to color 2. This distinction is relevant for discussing how humans decode encoded mappings (Section 2.2).

2.2 Assignment inference for decoding

When people decode encoded mappings, they use a process similar to solving an assignment problem, called *assignment inference* [30]. In assignment inference, people estimate the merit of different assignments based on association strengths between visual features and concepts, and determine the assignment that maximizes merit. However, unlike how computers solve assignment problems, human assignment inference is probabilistic rather than deterministic [30]. Overall, humans can produce reliable inferences, but their responses are noisy. This noise can be attributed to uncertainty in the color-concept associations that serve as input into the assignment problem. In [30], this uncertainty was built into models that were effective at predicting human responses.

Fig. 2 represents the noise in color-concept associations as distributions for each color-concept pairing. For each distribution, the mean corresponds to edge thickness in the bipartite graph. The variability is assumed to be normal. Assume each time a person does assignment inference, they draw a random value from these distributions to estimate merit for each edge. When the distributions are far apart (Fig. 2 left), random draws will consistently result in the same outcome of the assignment problem. However, when distributions overlap (Fig. 2 middle), random draws can result in different outcomes of the assignment problem, which results in more uncertainty in assignment inference [30]. This is the basis for our semantic distance metric in the present work (see Section 3.2.2).

Evidence suggests that people perform global assignment inference when interpreting the meanings of colors in information visualizations [30]. In some cases, the global solution conflicts with local solutions (Fig. 2, right). Such conflicts can result in people inferring that concepts are assigned to their most weakly associated colors, even when stronger candidate colors exist in the palette. Schloss et al. [30] first demonstrated this phenomenon using a recycling task: participants were presented with images of two colored bins, along with a word describing one "target" concept. There were two possible targets, paper and trash. When trash was the target and was presented with white and purple bins, participants were faced with a scenario like in Fig. 2, right. Trash was more strongly associated with white than with purple ($x_3 > x_4$), but so was paper ($x_1 > x_2$), and the association between paper and white was especially strong. Participants reliably discarded trash into the purple bin, even though trash was more strongly associated with white (analogous to blue in Fig. 2, right).

Schloss et al. [30] also assessed methods for calculating merit to generate the encoded mapping, one that maximized association strength (isolated merit function) and one that prioritized semantic discriminability over association strength (balanced merit function, although the term semantic discriminability was not used in [30]). Within both color palettes, responses were faster and more accurate when the target concept was more strongly associated with its correct color, but participants were more accurate for the balanced palette than the isolated palette. These results suggest interpretability increases with semantic discriminability. However, in addition to being more semantically discriminable, colors in the balanced palette were also further apart in CIELAB space (see Fig. S.1 in the Supplementary Material of the present paper). Thus, it is unclear if colors in the balanced color palette were easier to interpret because they were more semantically discriminable, more perceptually discriminable, or both.

3 Approach

In the present study, we assessed the independent effects of semantic discriminability and perceptual discriminability on participants' interpretations of bar graph data visualizations. The paradigm was the same as in Schloss et al. [30], but instead of interpreting colors of unlabeled trash and recycling bins, participants interpreted colors of unlabeled bars in a bar graph. On each trial, participants saw a graph containing two different colored bars, along with a target fruit concept described above the graph (Fig. 1, left). Their task was to indicate which colored bar, left or right, corresponded to the target fruit. Within each experiment, participants judged all pairwise combinations of eight colors for two fruits (cantaloupe and strawberry in Experiment 1, mango and watermelon in Experiment 2). The data from the present experiment and analysis code are at `github.com/SchlossVRL/semantic-discriminability`.

We used a previous dataset on color-concept associations from Rathore et al. [27] to select the colors for the present study (Section 3.1), define accuracy for the present tasks (Section 3.2.1), and quantify semantic distance (Section 3.2.2). In [27], participants rated association strengths between each of 12 fruits and each of 58 colors (UW-58 colors), uniformly sampled in CIELAB space ($\Delta E = 25$). This distance should be at least one noticeable difference [33, 34]. Further details on the methods of [27] are in Supplementary Material of the present paper.

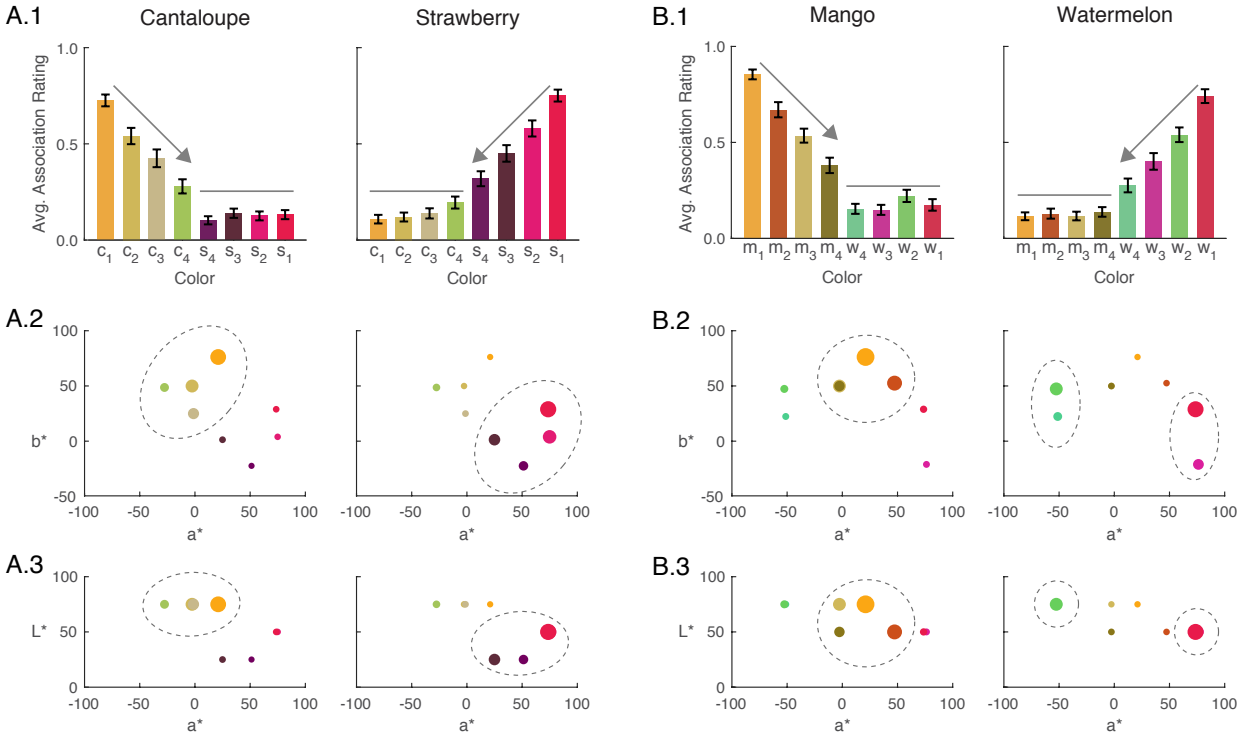

Figure 3. (A) Color-concept associations and CIELAB coordinates for colors tested in Experiment 1. (A.1) Mean association strength between each color with the concept labeled at the top of the column (error bars are standard errors of the means. Bar colors indicate the colors that were tested (see Table S.1 for coordinates). CIELAB coordinates are shown for each color on the (A.2) a*, b* plane and the a*, L* plane (A.3), with the size of the marks corresponding to association strength with the concept named at the top of the column (same data as in A.1). Dashed ovals enclose the four colors that were the relatively strong associates with the concept at the top of the column. (B) Same as A, but for the colors and concepts in Experiment 2. Each plot has eight points, one for each color, but in some cases fewer points are visible due to occlusion on the 2D plane.

3.1 Selecting colors and concepts

We chose the colors and fruit concepts using the mean color-concept association data from Rathore et al. [27] (see Table S.6 in the Supplementary Material) and color distances in CIELAB space (ΔE). In Experiment 1, we selected eight colors and two fruits to have the following properties. Four colors varied from moderately to strongly associated with the first fruit while being weakly associated with the second fruit (Fig. 3A.1, left). The other four colors varied from moderately to strongly associated with the second fruit while being weakly associated with the first fruit (Fig. 3A.1, right). We generated candidate colors and fruits using an optimization routine that enforced a minimum difference in association ratings for the four colors that varied and a weak association rating for the remaining four colors. This yielded a list of candidate palettes. We excluded palettes involving fruits that had colors in their names (blueberry and orange), and this led us to selecting cantaloupe and strawberry for our first experiment. The CIELAB coordinates of these colors are plotted on the a*, b* plane in Fig. 3A.2 and on the a*, L* plane in Fig. 3A.3. The size of the marks represent the association strengths shown in 3A.1. It can be seen that there are two separate clusters for "cantaloupe colors" and "strawberry colors" (indicated by the dashed ovals). In Experiment 2, we selected fruits and colors so they had the same color-concept association properties as in Experiment 1 (compare Fig. 3B.1 to 3A.1). However, unlike Experiment 1, the colors in Experiment 2 are no longer clustered in CIELAB space (Fig. 3B.2 and Fig. 3B.3); the "watermelon colors" were split on either side of the "mango colors". These properties enabled us to independently vary semantic distance and perceptual distance.

3.2 Quantifying metrics

3.2.1 Interpretability

We operationalized interpretability as the *accuracy* of decoding the encoded mapping. The bar graphs in this study were unlabeled, so there was no explicit encoded mapping from the perspective of the participants (i.e., no objectively correct answer). However, we can determine an optimal encoded mapping by solving an assignment problem for each pair of colors and concepts and use the solution to define the "correct" response. The input to the assignment problem was the set of association strengths between each color-concept pair (Fig. 3). Recall these data came from different participants than those in the present study. We solved the assignment problem for each pair of colors and concepts using the method described in Section 2.1.

3.2.2 Semantic discriminability

We operationalized semantic discriminability using a new metric, called *semantic distance* (ΔS). To build the intuition for semantic distance, consider semantic discriminability in the context of assignment problems, described in Section 2.1. Given color-concept association ratings, the solution to the assignment problem yields a deterministic assignment of colors to concepts. However, we want to distinguish between *robust* assignments (e.g., blueberry–blue and banana–yellow, which have high semantic discriminability), and *fragile* assignments (e.g., lemon–greenish-yellow and banana–orangish-yellow, which have low semantic discriminability since both fruits can be either color). In a robust assignment, we can expect all people to come to the same conclusion. But in a fragile assignment, people might disagree on which assignment is correct, and the same person might even respond differently when asked the same question again. We account for variability across individuals by assuming the association ratings between colors and concepts are normally distributed with a mean equal to the mean association rating and variance that is largest when the association rating is closest to the center of the rating scale.

We now define semantic distance in the case of two concepts and two colors and illustrate our definition in Fig. 4 using mango and watermelon as concepts and m_4 and w_4 as colors.

Given two colors and two concepts, there are two possible assignments of colors to concepts (indicated by black edges on the two bipartite graphs in Fig. 4). We define the semantic distance to be the absolute difference in the probabilities of each assignment being chosen by a random individual. Specifically, if x_1, x_2, x_3, x_4 are the

association ratings between colors and concepts (see Fig. 4), we let $\Delta x = (x_1 + x_4) - (x_2 + x_3)$. Note that if $\Delta x > 0$, concept M will be assigned with color m_4 and concept W will be assigned with color w_4. If $\Delta x < 0$, the alternative assignment will be made. We assume each x_i is normally distributed, with mean equal to $\bar{x}_i$, the mean association across all people for this color and concept, and standard deviation equal to $\sigma_i = 1.4 \cdot \bar{x}_i(1 - \bar{x}_i)$. This was found to be a good fit to the experimental data[1]. We define semantic distance as

$$\Delta S = \big|\, \text{Prob}(\Delta x > 0) - \text{Prob}(\Delta x < 0) \,\big|. \tag{1}$$

The probabilities in (1) can be computed by computing the z-score using the mean and standard deviations described above.

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

and $\text{Prob}(\Delta x < 0) = 1 - \text{Prob}(\Delta x > 0)$, where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal distribution. The relationship between the x_i and Δx is illustrated in Figure 4.

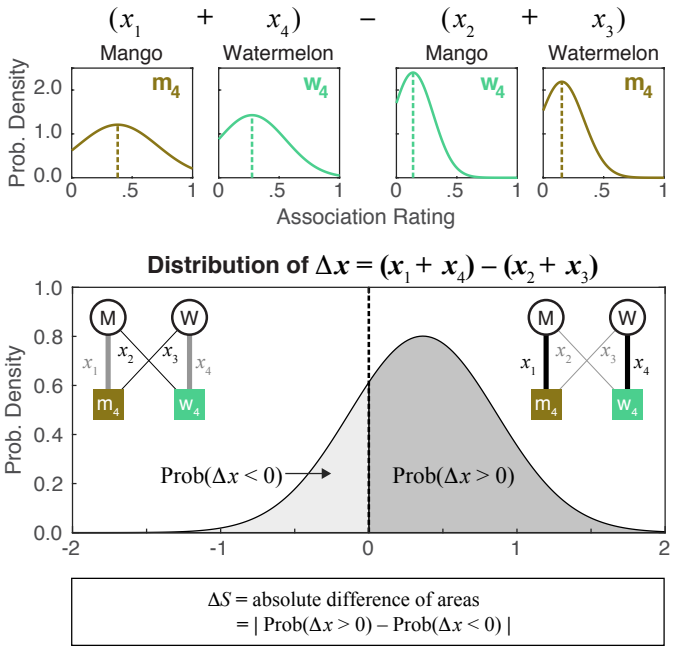

Figure 4. Illustration of semantic distance calculation using pairwise association ratings x_i between the colors m_4 and w_4 and the concepts Mango (M) and Watermelon (W). Distributions are normal distributions fit to individual participants' color-concept association ratings, with the dashed line showing the mean corresponding to the bars in Fig. 3B.1. The x_i are assumed to be normally distributed and combine to form Δx. When $\Delta x > 0$, the assignment M-m_4/W-w_4 is chosen and when $\Delta x < 0$, the alternative assignment M-w_4/W-m_4 is chosen. In the bipartite graphs, black/gray edges indicate the chosen/non-chosen assignment. Edge thickness indicates the mean color-concept association rating, but random draws can produce values above/below this mean, resulting in outcomes to left or right of zero. We define *semantic distance* ΔS as the absolute difference between these probabilities. We have $0 \le \Delta S \le 1$ and a larger ΔS indicates more certainty in the outcome of the assignment.

Fig. 2 left and center illustrate how semantic distance can vary while the assignment problem outcome remains constant. In both examples, concept A is assigned to color 1 and concept B is assigned to color 2, but semantic distance decreases between Fig. 2 left and center because the merit distributions overlap more. We predict that such decreases in semantic distance will make visualizations more difficult to interpret. In Fig. 2 right, the colors have a greater semantic distance than in Fig. 2 center, even though color 1 is more strongly associated than color 2 with both concept A and concept B. Thus, the scenario in Fig. 2 right should be more interpretable than Fig. 2 center. This example shows how colors can have a large semantic distance with respect to two concepts, even though neither color is strongly evocative of a particular concept within the set. Prior work has shown such cases are easily interpretable [30] (see Section 2.2).

Fig. 5A shows semantic distance for all 28 pairwise combinations of 8 colors tested in Experiment 1 (see figure caption for details on how to interpret this plot). There is only one plot for both cantaloupe and strawberry because semantic distance for a given set of colors and concepts is symmetric. That is, the distance between colors c_1 and s_1 is the same, regardless of whether the target is cantaloupe or strawberry. When cantaloupe colors are paired with other cantaloupe colors (on curves labeled c_1, c_2, c_3), semantic distance increases as the difference in association strength increases (i.e., distance on the x-axis), but then levels off once reaching strawberry colors because all strawberry colors are similarly weakly associated with cantaloupe. To see the analogous pattern for strawberry colors, it is necessary to compare the heights of data points at each x-axis position. For example, looking at s_1 on the x-axis, semantic distance steadily increases for pairings with other strawberry colors as association strength difference increases (s_2 to s_4), and then levels off when reaching the four cantaloupe colors.

Fig. 5B shows semantic distance for the colors and concepts in Experiment 2. Given that the pattern of association strengths across colors in Experiment 2 (Fig. 3B.1) was similar to Experiment 1 (Fig. 3A.1), the pattern of semantic distances were strongly correlated between the two experiments ($r(26) = .99, p < .001$).

3.2.3 Perceptual discriminability

We operationalized perceptual discriminability as *perceptual distance* (ΔE) in CIELAB color space, as in previous visualization research [33, 34]. Fig. 5C shows perceptual distances in Experiment 1 and Fig. 5D shows perceptual distances in Experiment 2, plotted in the same manner as semantic distance. In Experiment 1, perceptual distance (Fig. 5C) deviated from semantic distance (Fig. 5A) but the two variables were still correlated ($r(26) = .71, p < .001$). In Experiment 2, perceptual distance (Fig. 5D) and semantic distance (Fig. 5B) were uncorrelated ($r(26) = .02, p = .920$). Perceptual distances in Experiment 1 and Experiment 2 were also uncorrelated ($r(26) = .08, p = .673$)

4 Experiment 1

This experiment tested for independent effects of semantic and perceptual distance on interpretability, using the colors in Fig. 3A. Although semantic distance and perceptual distance were correlated, we could test for independent effects of each factor using regression analyses.

4.1 Methods

Participants. 36 undergraduates (mean age = 18.3, 25 females, 11 males) participated for credit in Introductory Psychology. All had normal color vision (screened with [10]), and gave informed consent. The UW–Madison IRB approved the protocol for this study.

Design and Displays. Participants were presented with bar graphs showing fictitious data about preferences for two different fruits (Fig. 1, left). Each graph had two colored bars, one for each fruit. The bars were two different colors, determined by all 28 pairwise combinations of eight colors (Fig. 3A, Table S.1 in the Supplementary Material). The bars were 50 pixels wide (2.4 cm wide) and varied in height. Each trial contained a taller and shorter bar with base heights of 150 and 100 pixels (5.1 and 3.7 cm), respectively. Bar heights were randomly, and independently, adjusted around their base height by +/- 5 pixels (.2 cm) on each trial. The side of the graph containing the taller bar was left/right balanced. The x and y axes of the graph were 200 and 250 pixels long respectively. The y-axis was labeled as "Preference" (font size 14) and the x-axis and bars were unlabeled. The target fruit for

[1]Many other choices could be made here, by picking other functions that have a similar qualitative shape (i.e., zero standard deviation when $\bar{x} = 0$ or 1 and maximum standard deviation when $\bar{x} = 0.5$). We experimented with other options and found our results to be robust with respect to the choice of function.

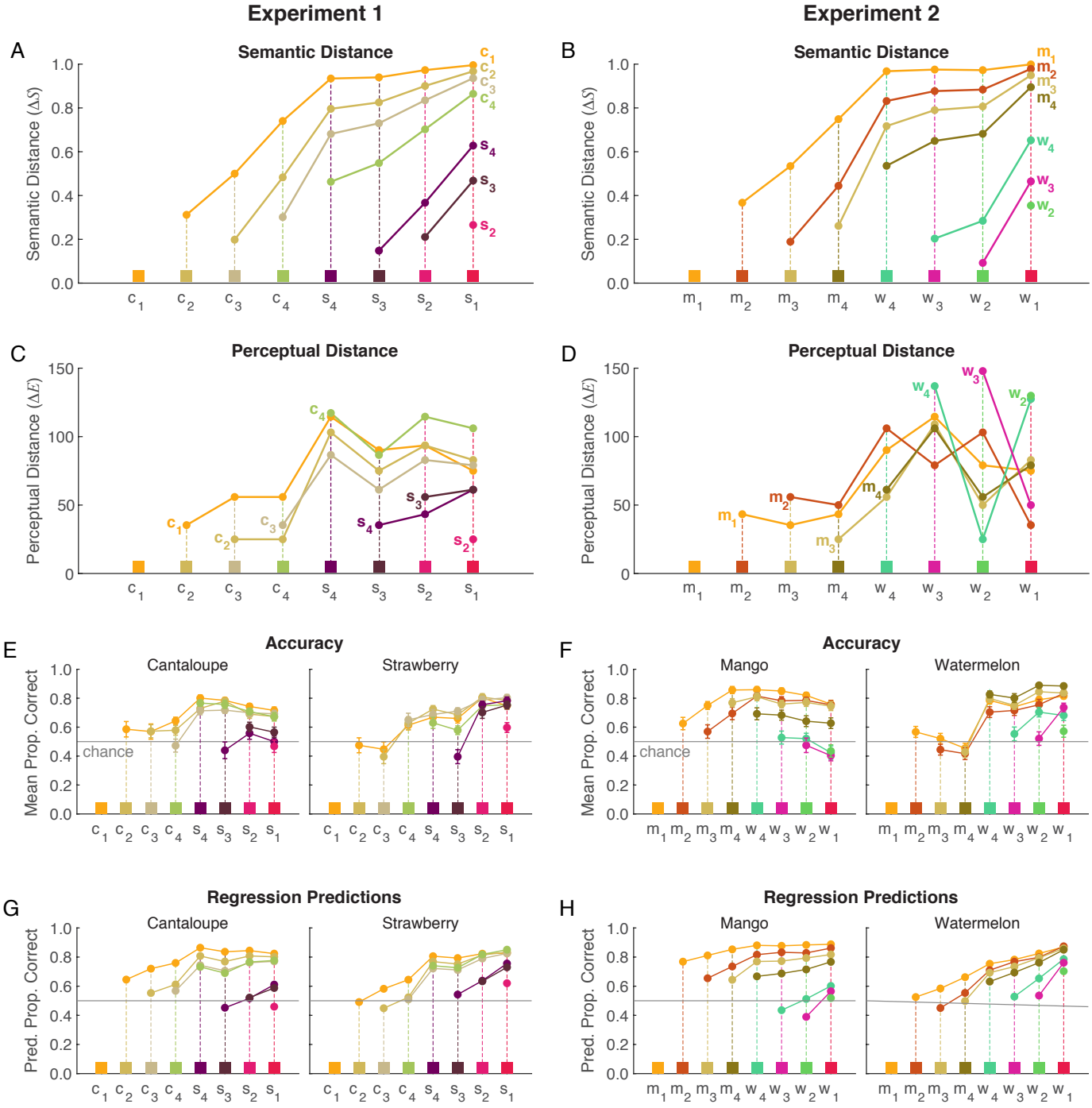

Figure 5. The left column (A,C,E,G) refers to Experiment 1 (cantaloupe and strawberry) and the right column (B,D,F,H) shows the analogous data for Experiment 2 (mango and watermelon). We will describe the left column. The top row shows semantic distances (ΔS) for all pairs of colors. Since ΔS for a pair of colors is defined in the context of both target fruits, we can represent the data in a single plot. The plot contains 28 points, one for each pair of distinct colors from the set of 8 colors. Each of the 28 points is identified with a pair of colors as follows. The color of the point itself identifies the first color, and the vertical dashed line crossing that point leads to a label on the x-axis, identifying the second color. Thus, all points connected by dashed lines share a common color, and likewise for the solid lines. The colored squares along the x-axis and associated labels also serve as a legend for the mark colors. In (A), colors c_1 to c_4 are the colors most strongly associated with cantaloupe and s_1 to s_4 are the colors most associate with strawberry (lower subscripts are more strongly associated with the fruit indicated by the letter, see also Fig. 3). The second row shows perceptual distance (ΔE), plotted in the same manner as semantic distance. The third row shows mean proportion of correct responses plotted separately for each target concept because each target was assessed independently. Error bars represent standard errors of the means using the Cousineau [5] adjustment to account for overall differences at the subject level. The bottom row shows predicted response accuracy using regression equations from Table 1.

a given trial was displayed as text positioned above the graph (20 pt font), centered on the x-axis. Thus, the full experiment design included 2 target concepts (cantaloupe or strawberry) $\times$ 28 color pairs $\times$ 2 positions of the colors within each pair (left or right) $\times$ 2 taller bar sides (left or right) $\times$ 3 repetitions, producing 672 trials.

The displays were generated and presented using Presentation (`www.neurobs.com`). The monitor was a 24.1 in ASUS ProArt PA249Q monitor (1920×1200 resolution), viewed from about 60 cm. The background was gray (CIE Illuminant D65, x = .3127, y = .3290, Y = 10 cd/m^2). We used a white point of D65, luminance = 100 cd/m^2) to convert between CIELAB and CIE 1931 xyY coordinates. We used a Photo Research PR-655 SpectraScan spectroradiometer to characterize the monitor and verify accurate presentation of the colors. The deviance between the measured and target colors in CIE 1931 xyY coordinates was $< .01$ for x and y, and < 1 cd/m^2 for Y.

Procedure. The participants were told that they would be presented with a series of bar graphs, each showing a different person's preferences for two fruits, cantaloupe and strawberry. Within each graph, one bar would represent cantaloupe and the other bar would represent strawberry. The bars would have different colors, but would not be labeled. Above the graph, participants would see the name of one of the two fruits, cantaloupe or strawberry. Their task was to decide which bar corresponded to the fruit described above the graph and to respond by pressing the corresponding arrow key (left or right). Participants were reminded that the bars would not be labeled, and were asked to use their intuition about which bar color corresponded to the fruit described.

Participants then completed five practice trials drawn at random from all possible trials. They then completed the 672 test trials, presented in a blocked randomized design (three blocks to accommodate three repetitions). Each block included all combinations of targets, color pairs, color positions, and bar height positions, presented in a random order. Participants received a break after each set of 28 trials. Stimuli remained on the screen until participants responded, and trials were separated by a 500-ms inter-trial interval. We recorded which color was chosen and the response time (RT) to make the choice on each trial.

4.2 Results and Discussion

We first present results on accuracy, where "correct" was defined as the solution to the assignment problem for both possible targets and the two colors on a given trial (see Sections 2.1 and 3.2.1). We then present results on RTs, which can be interpreted as how difficult it was to make the decision on each trial, regardless of accuracy.

Accuracy. Fig. 5E shows mean accuracy for each color pair when the target was cantaloupe (left) or strawberry (right). To obtain these means, we first calculated the proportion of correct trials for each participant, for each target (cantaloupe or strawberry) and each pair of colors (all 28 combinations of 8 colors). This proportion included 12 data points (2 left/right positions of the colors within each pair $\times$ 2 positions of the taller bar within the bar graph $\times$ 3 repetitions). We then calculated the mean for each target and color pair across participants.

The correct color for each target and each color pair is indicated by the color positions on the x-axes of Fig. 5E. Within each color pair, the color toward the left on the x-axis was correct for cantaloupe, and the color toward the right was correct for strawberry. For example, given c_2 and c_4, c_2 was correct for cantaloupe and c_4 was correct for strawberry.

We first highlight three key observations in Fig. 5E. First, most of the responses were well above chance. This means that participants could reliably decode our encoded mappings, even though there was no legend. Second, there is systematic variability in response accuracy across color pairs. This provides further support that human assignment inference is probabilistic, not deterministic. Recall that if a computer were solving assignment problems in our task, the outcome would be deterministic—responses would all be at 1.0 regardless of whether the assignment problem was robust or fragile (see Sections 2.1 and 3.2.2).

Third, the pattern of accuracies resembles aspects of semantic distance (Fig. 5A) and perceptual distance (Fig. 5C). Like both predictors, accuracy tends to be greater when pairs include one cantaloupe color (c_1 - c_4) and one strawberry color (s_1 - c_4) ("between-concept pairs"), especially for cantaloupe. The predictors differ in that semantic distance increased monotonically from left to right in 5A, whereas perceptual distance is non-monotonic based on how we selected the colors. Perceptual distance is flatter among "within-concept" color pairs where colors are perceptually similar (among c_1 - c_4 and among s_1 - s_4), and fluctuates systematically across between-concept color pairs (Fig 5C). Accuracy resembles the flatness of perceptual distance for within-concept pairs where colors were most perceptually similar (especially for cantaloupe), but accuracy resembles the smoothness of semantic distance for between-concept pairs where colors were most perceptually distinct.

Table 1. Mixed-effect logistic regression models of accuracy in Experiment 1 (Acc 1.1, Acc 1.2) and Experiment 2 (Acc 2.1, Acc 2.2).

Model	Factor	β	SE	z	p
Acc 1.1	Intercept	0.89	0.14	6.37	$<.001$
	PercDist	0.22	0.06	3.58	$<.001$
	SemDist	0.34	0.07	4.96	$<.001$
Acc 1.2	Intercept	0.91	0.14	6.38	$<.001$
	PercDist	0.26	0.06	4.11	$<.001$
	SemDist	0.23	0.07	3.05	.002
	Assoc	0.23	0.05	4.77	$<.001$
Acc 2.1	Intercept	0.97	0.12	8.36	$<.001$
	PercDist	-0.06	0.03	-1.90	.057
	SemDist	0.55	0.06	9.24	$<.001$
Acc 2.2	Intercept	1.00	0.12	8.45	$<.001$
	PercDist	-0.06	0.03	-1.84	.066
	SemDist	0.41	0.06	6.86	$<.001$
	Assoc	0.37	0.05	7.95	$<.001$

To test for independent effects of perceptual and semantic distance on accuracy, we used a mixed effect logistic regression (R version 4.0.2, lme4 1.1-23). The dependent measure was accuracy on each trial for each participant (1 = correct, 0 = incorrect). We included fixed effects for semantic distance and perceptual distance, and random slopes and intercepts for subjects within each fixed effect. We used z-scores of the predictors in all models to center them and put them on similar scales. As shown in Table 1 (Model Acc 1.1), accuracy significantly increased with increased semantic distance and perceptual distance.

Recall that both semantic and perceptual distance are symmetric, they are defined with respect to a given color pair, irrespective of the target. Thus, based on these factors alone, we would predict that the pattern of responses for both targets would be the same. However, as shown in Fig. 5E, there are systematic asymmetries. In particular, note how accuracy among pairs including the strawberry colors was greater for strawberry targets than cantaloupe targets. To fully capture this pattern of data, it is necessary to add another predictor that accounts for differences depending on the target concept.

Thus, we repeated the same model but added a new factor that could capture target-specific responses: association strength between the target and the correct color. This factor was previously shown to predict accuracy and RTs for similar data [30]. As shown in Table 1 (model Acc 1.2), association strength significantly predicted accuracy, and the previous two factors were still significant. Therefore, accuracy increased with semantic distance, perceptual distance, and association strength between the target and the correct color. Fig. 5G shows the predicted data using weights from model Acc 1.2 in Table 1. The model predictions and data for all 28 color pairs $\times$ 2 concepts are strongly correlated ($r(54) = .82, p < .001$).

We also examined the relation between predictors in the model. Across the 28 color pairs for each of the two targets, association strength between the target and correct color was moderately correlated with semantic distance ($r(54) = .43, p < .001$) and not significantly correlated with perceptual distance ($r(54) = .21, p = .123$).

Response time. Fig. 6A shows mean RTs for each color pair, obtained by first calculating the median RT across all 12 trials for each target and color pair for each participant, and then calculating the mean over participants. Treating RTs this way avoids effects of outliers

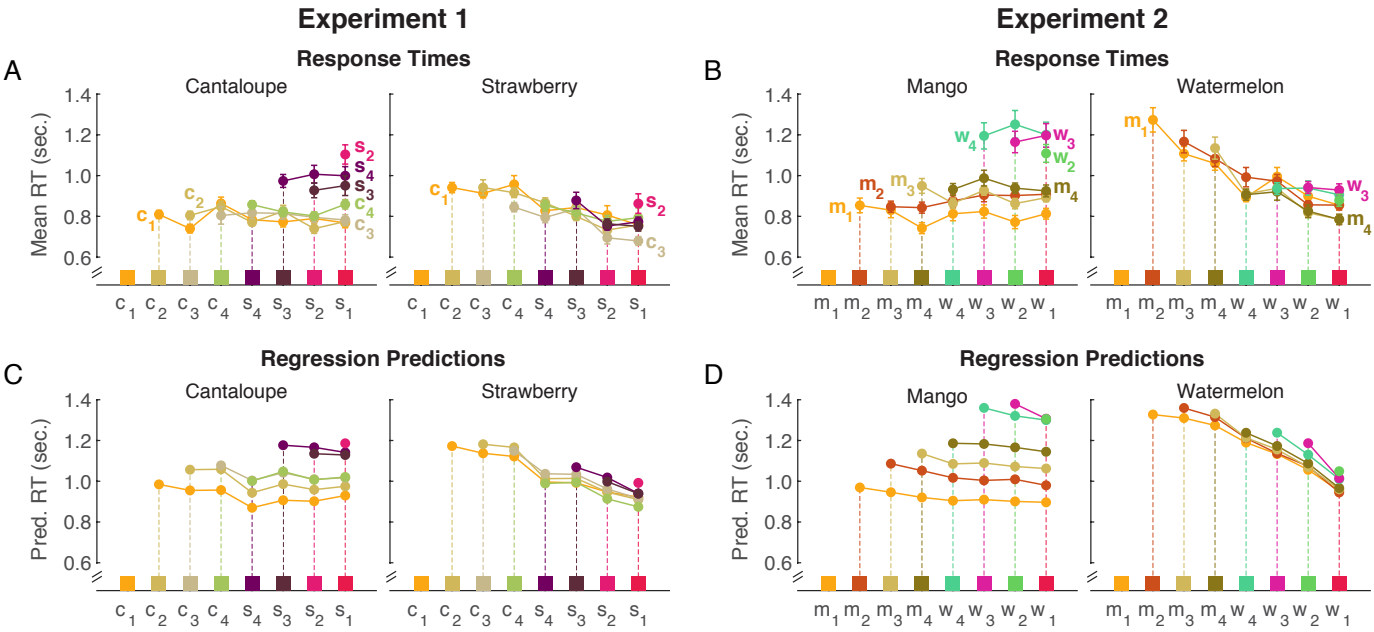

Figure 6. The top row shows mean RTs for (A) Experiment 1 and (B) Experiment 2, plotted in the same manner as in Fig. 5. Error bars represent standard errors of the means, using the Cousineau [5] adjustment to account for overall differences at the subject level. The bottom row shows regression predictions for RTs using the model with all three predictors in Table 2 in (C) Experiment 1 and (D) Experiment 2.

without excluding trials [26]. Mean RTs were negatively correlated with mean accuracy ($r(54) = -.71, p < .001$), such that participants responded more quickly for color pairs that facilitate accuracy.

We analyzed the RT data using linear-mixed effect models with the same predictors as for accuracy. The model results are in Table 2. When perceptual distance and semantic distance were the only fixed effects (model RT 1.1), neither predictor explained significant variance. When association strength was added to the model, it explained significant variance, as did perceptual distance (model RT 1.2). Thus, RTs were faster when association strength was stronger and when perceptual distance was larger. Fig. 6C shows the model prediction using the weights from model RT 1.2 in Table 2. Model predictions were strongly correlated with mean RTs ($r(54) = .82, p < .001$).

Table 2. Linear mixed-effects regression models of RT in Experiment 1 (RT 1.1, RT 1.2) and Experiment 2 (RT 2.1, RT 2.2).

Model	Factor	β	SE	*df*	*t*	*p*
RT 1.1	Interept	1017.7	53.2	35.0	19.1	<.001
	PercDist	-29.3	17.1	121.4	-1.7	.088
	SemDist	-37.2	20.6	39.0	-1.8	.078
RT 1.2	Interept	1017.7	53.2	35.0	19.1	<.001
	PercDist	-42.8	17.0	156.9	-2.5	.013
	SemDist	1.9	19.6	78.6	0.1	.924
	Assoc	-68.3	15.9	48.0	-4.3	<.001
RT 2.1	Intercept	1121.5	62.3	35.0	18.0	<.001
	PercDist	12.1	8.0	35.0	1.5	.139
	SemDist	-86.4	15.1	35.0	-5.7	<.001
RT 2.2	Intercept	1121.5	62.3	35.0	18.0	<.001
	PercDist	9.0	7.9	35.0	1.1	.260
	SemDist	-36.0	9.9	35.0	-3.6	<.001
	Assoc	-120.6	20.0	35.0	-6.0	<.001

In summary, semantic distance and perceptual distance both independently contributed to interpretability. However, the "cantaloupe colors" and "strawberry colors" were clustered in different parts of color space, (Fig. 3A.2-3), so conflicts between semantic and perceptual distance were only minor. In Experiment 2, we address the question of how these two factors would contribute to interpretability if they were overall decorrelated and included examples of large conflicts.

5 Experiment 2

This experiment tested for independent effects of semantic and perceptual distance when these two factors were uncorrelated. The pattern of association strengths was similar to Experiment 1 (Fig. 3A.1 and B.1), so the pattern of semantic distances were also similar (Fig. 5A and 5B). However, the relative locus of colors in CIELAB space was different. In Experiment 1, the strong associates for each concept clustered together (Fig. 3A.2-A.3), but in Experiment 2, the four "watermelon colors" were split on either side of the "mango colors" along the a* axis (Fig. 3B.2-B.3). Thus, for watermelon, the most semantically similar colors were furthest in color distance.

5.1 Methods

36 undergraduates (mean age = 19.4, 18 females, 18 males) participated for credit in Introductory Psychology. All had normal color vision (screened with [10]), and gave informed consent. The design, displays, and procedure were the same as Experiment 1, except we tested the colors and fruits in Fig. 3B (Table S.1).

5.2 Results and discussion

The colors and fruits in Experiment 1 and 2 differed in that their patterns of perceptual distances were uncorrelated ($r(26) = .08, p = .673$) but their patterns of semantic distances were almost perfectly correlated ($r(26) = .99, p < .001$). Thus, if the patterns of data in Experiment 2 are similar to Experiment 1, they can be attributed to their similarities in semantic distance. Fig. 5F shows the mean accuracy data and Fig. 6B shows the mean RTs, calculated in the same manner as in Experiment 1. Indeed, there were significant correlations between the patterns of accuracy ($r(54) = .66, p < .001$) and RT ($r(54) = .79, p < .001$) between the two experiments.

Accuracy. We analyzed accuracy using the same mixed-effect logistic regression models as in Experiment 1. The first model including perceptual distance and semantic distance showed that semantic distance significantly predicted accuracy (Table 1, model Acc 2.1). The effect of perceptual distance was marginal, but it was in the opposite direction from Experiment 1. That is, accuracy tended to increase for

more perceptually *similar* colors, probably because colors that were perceptually similar (e.g., w_1 (red) and m_2 (dark orange)) were semantically different, whereas colors that were perceptually distant were semantically similar (e.g., w_1 (red) and w_2 (green)) (Fig. 1).

When we added association strength between the target and correct color into the model, association strength was a significant predictor, as was semantic distance (Table 1, model Acc 2.2). The effect of perceptual distance was still marginal, again in the opposite direction (more perceptually different tended to result in reduced accuracy). Fig. 5H shows the predicted accuracy based on the regression weights in model Acc 2.2. The model predictions strongly correlated with the mean accuracy data in Fig. 5F ($r(54) = .83, p < .001$). In this experiment, association strength between the target and correct color was again moderately correlated with semantic distance ($r(54) = .42, p = .001$) and not significantly correlated with perceptual distance ($r(54) = -.02, p = .900$).

Response time. As in Experiment 1, RT and accuracy were negatively correlated ($r(54) = -.82, p < .001$), indicating it was easier to make decisions for color pairs that facilitated accuracy. We analyzed RTs using the same linear mixed-effect models from Experiment 1. The first model including only perceptual distance and semantic distance showed a significant effect of semantic distance and no effect of perceptual distance (Table 2, model RT 2.1). Adding association strength between the target and the correct color (model RT 2.2) resulted in significant effects of association strength and semantic distance but still not perceptual distance. Fig. 6D shows the predicted RTs based on the regression weights in model RT 2.2. The model predictions strongly correlated with the mean RTs in Fig. 6B ($r(54) = .88, p < .001$).

In summary, semantic distance dominated interpretability when these two factors were uncorrelated overall. Perceptual distance had a marginal effect, but it was in the opposite direction from what might be expected (i.e., smaller perceptual distances tended to be more interpretable). This was because the stimulus set included cases with strong conflicts, such that large semantic distances amounted to small perceptual distances (especially for watermelon), and under such conflicts greater semantic distance resulted in greater interpretability.

6 General Discussion and Conclusion

In this study we tested whether people's ability to interpret color palettes in information visualizations depended on semantic distance, independent of perceptual distance. The results of both experiments demonstrated that increasing semantic distance improved interpretability, independent of variation in perceptual distance. In Experiment 1, we selected colors such that perceptual and semantic distance co-varied: the four colors that were most strongly associated with cantaloupe were clustered separately from the colors most strongly associated with strawberry. Under these conditions, both semantic and perceptual distance independently contributed to increased interpretability. In Experiment 2, we selected the colors in a way that decoupled perceptual and semantic distance: the four colors that were most strongly associated with mango were between the colors most strongly associated with watermelon on the a* plane of CIELAB space. Across all color pairs in Experiment 2, perceptual distance and semantic distance were uncorrelated, but there were cases in which these two factors were in direct conflict (Fig. 1). In this experiment, accuracy and RT both improved with increased semantic distance, with no significant effects of perceptual distance. The results of this study suggest that it may be worth relaxing constraints on perceptual distance in favor of maximizing semantic distance to create interpretable color palettes.

We studied colors that were distant enough ($\Delta E \geq 25$) to be noticeably different [33, 34], but we expected that perceptual distance would play a larger role if distances were smaller. However, if colors were no longer perceptually discriminable, they would also no longer be semantically discriminable. Thus, thresholding at some degree of semantic distance may be sufficient to ensure both perceptual and semantic discriminability. Certainly, there is some lower threshold at which perceptual and semantic discriminability would be too small for interpretability, but there also may be an upper threshold at which further increasing perceptual or semantic discriminability would have no further benefit. Substantial work has investigated lower thresholds for perceptual discriminability for information visualizations [33, 34], but future work is needed to understand thresholds for semantic discriminability. Moreover, as in prior visualization work [33, 34] we used ΔE as our perceptual distance metric, but future work could evaluate whether different perceptual distance metrics (e.g., CIEDE2000) are better at predicting interpretability.

As part of this study, we developed semantic distance, ΔS, as a metric to quantify semantic discriminability between pairs of colors and concepts. Semantic distance is the absolute difference in the probabilities that a random observer will make each of the two possible assignments, where the randomness is due to inherent variability in association strengths across individuals. Quantifying semantic discriminability becomes more difficult when there are more than two colors or two concepts because there are more than two possible assignments. Solving assignment problems becomes more complicated in this case, and we cannot write a simple formula as in (2) to compute assignment probabilities.[2] Possible approaches for quantifying semantic discriminability for more than two colors and concepts could involve obtaining a distribution over possible assignments (e.g., via Monte Carlo simulation), resulting from uncertainty in color-concept association ratings and applying one of many possible metrics. For example, if we used *entropy*, maximum entropy would correspond to the fragile case of all assignments having equal probability. Conversely, minimum entropy would correspond to the robust case of one assignment having a very high probability and all other assignments having near-zero probability.

In this study, we used mean human color-concept association ratings to quantify association strengths. However, efficient automated approaches exist for estimating color-concept associations using images [19–21, 27] and natural language databases [13, 31]. Different methods can be used to extract colors from images, but evidence suggests methods that leverage perceptual dimensions of color and cognitive representations of color categories are best for estimating human color-concept associations [27]. Such estimates, combined with an appropriate method for quantifying variance in the sample images, could be used as input to calculate semantic distance.

The results of this study can help with designing interpretable color palettes, but interpretability is only one of the many goals in color palette design. Other priorities might include helping observers (1) locate a target in visual search [9, 11, 14, 35, 38], (2) estimate the area of colored regions [1, 8], (3) refer to the colors easily by name [15], (4) appreciate the visualization aesthetically [8], or (5) obtain an affective impression from the overall palette [2]. Different design properties are relevant for these different priorities. For example, the ability to estimate the relative area occupied by colored regions increases with perceptual distance between colors, but aesthetic preferences for those same visualizations decreases with perceptual distance [8]. Extensive work is needed to understand how to navigate such trade-offs in palette design, depending on the priorities and format of a given visualization [2, 8, 19, 33]. The present work provides a step in that direction by showing that maximizing perceptual distance is not necessary for creating interpretable color palettes, leaving room for maximizing the other factors that contribute to effective palette design.

Acknowledgments

We thank Ragini Rathore, Kevin Lande, Kushin Mukherjee, Melissa Schoenlein, Anna Bartel, Brian Yin, and Joris Roos for their feedback on this work, and Shannon Sibrel, Autumn Wickman, Marin Murack, Yuke Liang, Charles Goldring, and Brianne Sherman for help with data collection. This work was supported by the UW–Madison Office of the Vice Chancellor for Research and Graduate Education and Wisconsin Alumni Research Foundation.

[2] Generally, solving a large assignment problem (many colors and concepts) cannot be reduced to solving a sequence of smaller two-color two-concept assignment problems. For example, the colors (red,yellow) and concepts (apple,banana) have a straightforward assignment, but if we add another color and concept: (red,yellow,green) and (apple,banana,strawberry), then apple switches from red to green due to the presence of strawberry.

REFERENCES

[1] D. Albers, M. Correll, and M. Gleicher. Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 551–560, 2014.

[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] C. A. Brewer. Color use guidelines for mapping and visualization. In A. M. MacEachren and D. R. F. Taylor, editors, *Visualization in Modern Cartography*, pages 123–148. Elsevier Science Inc., Tarrytown, 1994.

[4] W. S. Cleveland and R. McGill. Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society: Series A (General)*, 150(3):192–210, 1987.

[5] D. Cousineau et al. Confidence intervals in within-subject designs: A simpler solution to loftus and masson's method. *Tutorials in Quantitative Methods for Psychology*, 1(1):42–45, 2005.

[6] D. J. Cuff. Colour on temperature maps. *The Cartographic Journal*, 10(1):17–21, 1973.

[7] D. Gentner and C. Toupin. Systematicity and surface similarity in the development of analogy. *Cognitive Science*, 10(3):277–300, 1986.

[8] C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):521–530, 2017.

[9] C. C. Gramazio, K. B. Schloss, and D. H. Laidlaw. The relation between visualization size, grouping, and user performance. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1953–1962, 2014.

[10] L. H. Hardy, G. Rand, M. C. Rittler, J. Neitz, and J. Bailey. *HRR Pseudoisochromatic Plates*. Richmond Products, 2002.

[11] S. Haroz and D. Whitney. How capacity limits of attention influence information visualization effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2402–2410, 2012.

[12] M. Harrower and C. A. Brewer. ColorBrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

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

[14] C. G. Healey. Choosing effective colours for data visualization. In *Proceedings of the 7th Conference on Visualization'96*, pages 263–ff. IEEE Computer Society Press, 1996.

[15] J. Heer and M. Stone. Color naming models for color selection, image editing and palette design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1007–1016. ACM, 2012.

[16] M. Hegarty. The cognitive science of visual-spatial displays: Implications for design. *Topics in Cognitive Science*, 3(3):446–474, 2011.

[17] S. M. Kosslyn and S. M. Kosslyn. *Graph Design for the Eye and Mind*. OUP USA, 2006.

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

[19] 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. Wiley Online Library, 2013.

[20] 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.

[21] 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.

[22] M. McGranaghan. Ordering choropleth map symbols: The effect of background. *The American Cartographer*, 16(4):279–285, 1989.

[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] D. Norman. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books (AZ), 2013.

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

[26] R. Ratcliff. Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114(3):510, 1993.

[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] I. T. Ruginski, A. P. Boone, L. M. Padilla, L. Liu, N. Heydari, H. S. Kramer, M. Hegarty, W. B. Thompson, D. H. House, and S. H. Creem-Regehr. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, 16(2):154–172, 2016.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] D. A. Szafir. Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):392–401, 2018.

[35] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[36] B. Tversky. Visualizing thought. *Topics in Cognitive Science*, pages 499 – 535, 2011.

[37] B. Tversky, J. B. Morrison, and M. Betrancourt. Animation: can it facilitate? *International Journal of Human-Computer Studies*, 57(4):247–262, 2002.

[38] J. M. Wolfe and T. S. Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):1–8, 2017.

[39] M. Wood. Visual perception and map design. *The Cartographic Journal*, 5(1):54–64, 1968.

[40] C. Xiong, L. van Weelden, and S. Franconeri. The curse of knowledge in visual data communication. *IEEE Transactions on Visualization and Computer Graphics*, 2019.

[41] J. Zacks and B. Tversky. Bars and lines: A study of graphic communication. *Memory & Cognition*, 27(6):1073–1079, 1999.