Unifying Effects of Direct and Relational Associations for Visual Communication

Melissa A. Schoenlein, Johnny Campos, Kevin J. Lande, Laurent Lessard, and Karen B. Schloss

Abstract—People have expectations about how colors map to concepts in visualizations, and they are better at interpreting visualizations that match their expectations. Traditionally, studies on these expectations (inferred mappings) distinguished distinct factors relevant for visualizations of categorical vs. continuous information. Studies on categorical information focused on direct associations (e.g., mangoes are associated with yellow) whereas studies on continuous information focused on relational associations (e.g., darker colors map to larger quantities: dark-is-more bias). We unite these two areas within a single framework of assignment inference. Assignment inference is the process by which people infer mappings between perceptual features and concepts represented in encoding systems. Observers infer globally optimal assignments by maximizing the “merit,” or “goodness,” of each possible assignment. Previous work on assignment inference focused on visualizations of categorical information. We extend this approach to visualizations of continuous data by (a) broadening the notion of merit to include relational associations and (b) developing a method for combining multiple (sometimes conflicting) sources of merit to predict people’s inferred mappings. We developed and tested our model on data from experiments in which participants interpreted colormap data visualizations, representing fictitious data about environmental concepts (sunshine, shade, wild fire, ocean water, glacial ice). We found both direct and relational associations contribute independently to inferred mappings. These results can be used to optimize visualization design to facilitate visual communication.

Index Terms—Visual reasoning, information visualization, colormap data visualizations, visual encoding, color cognition

1 INTRODUCTION

Imagine you are interpreting a bar chart and need to infer which colors map to which quantities represented in the chart. Now, imagine instead interpreting a colormap data visualization1 and you need to infer which colors map to which quantities represented in the colormap.

Traditionally, researchers studying the role of color semantics for visual communication have treated these cases as two distinct problems. One involves mapping colors to different categories in categorical information [18, 22, 36, 38, 41] and the other involves mapping gradations of color to gradations of quantity in continuous data [8, 21, 35, 43]. In both cases, a key goal is to understand people’s expectations about the mappings between colors and concepts in visualizations (called inferred mappings) because visualizations designed to match people’s expectations are easier to interpret [14, 18, 22, 26, 35, 36, 38, 43, 50, 51].

Studies on visualizations of categorical information focus on direct associations—the degree to which each color is associated with each concept represented in the visualization. Methods have been developed to use direct associations to optimize mappings between discrete colors and concepts to facilitate visualization interpretability [18, 22, 36, 38, 41].

Studies on visualizations of continuous data focus on relational associations—correspondences between relational properties of visual features and relational properties of concepts. For example, observers have a dark-is-more bias, inferring that darker colors map to larger quantities [4, 8, 21, 35, 43]. This bias is relational because it depends of colors to quantities (e.g., weather maps, neuroimaging maps, correlation matrices). “Color scale” refers to the color gradient used to construct a colormap.

---

1Literature on visualizing continuous data using color has inconsistent terminology. In this paper, “colormap” refers to a visualization that maps gradations of colors to quantities (e.g., weather maps, neuroimaging maps, correlation matrices). “Color scale” refers to the color gradient used to construct a colormap.
on relative lightness, rather than particular colors in visualizations. Although empirical studies of colormaps have focused on relational associations and explicitly tried to avoid potential effects of direct associations [21, 35, 43], direct associations likely play an important role (see Samsel et al.’s [34] intuitive colormaps for environmental visualizations).

In this paper, we aim to unite the study of direct and relational associations under a single framework of *assignment inference*. Assignment inference is the process by which people infer mappings among visual features and concepts in visual encoding systems [38]. Previous work on assignment inference focused on visualizations of categorical information, showing that observers infer optimal assignments (i.e., mappings) that maximize the total “goodness” of each possible color-concept pair [22, 36, 38]. This “goodness” is called merit.

We propose that assignment inference also governs inferences about the meanings of colors in visualizations of continuous data. In testing this possibility, our work makes the following contributions: (1) We broaden the notion of “merit” in assignment inference to include relational associations, and show that both relational and direct associations influence inferred mappings for colormap visualizations. (2) We develop a method for combining multiple (sometimes conflicting) sources of merit for simulating assignment inference, and show that our method effectively predicts inferred mappings for colormap visualizations.

### 2 Background

In this section, we review previous work on color semantics in information visualization. Following tradition, we discuss effects of direct associations for visualizations of categorical information and relational associations for visualizations of continuous data. We will unite these two areas in Section 3 on our approach in the present study.

#### 2.1 Direct associations and assignment inference

Direct associations (a.k.a. color-concept associations) are the degree to which a color is associated with a concept. They are estimated using various measures, including human judgments [1, 15, 24, 27, 31, 36, 38, 39, 49], image statistics [18, 19, 31, 41], and language corpora [13, 41]. Although direct associations influence inferred mappings between colors and concepts in visualizations of categorical information [18, 22, 36, 38], direct associations and inferred mappings are not the same thing. Cases arise in which people infer that a concept maps to a weakly associated color, even when there are more strongly associated colors in a visualization. This distinction is shown in Fig. 2A. The bipartite graph (left) represents association strengths between each of two colors (purple and white) and each of two concepts (trash (T) and paper (P)) in an encoding system for recycling bins [38]. The thickness of the edges connecting colors and concepts represents direct association strength (thicker indicates stronger association). Trash is more associated with white than with purple (thicker edges). Yet, when asked which colored bin is for trash (Fig. 2A right), people choose purple. Why?

Evidence suggests the reason is that people approach this problem using assignment inference, a process that considers all colors and concepts in the scope of the encoding system [38]. Assignment inference is analogous to solving an assignment problem in optimization [23]. In Fig. 2A, the scope of the encoding system includes trash and paper, even though paper was not relevant on this particular trial. Assignment inference does not simply assign a color to the concept with the strongest merit (for now, think of merit as direct association strength). Instead, the process selects the combination of color-concept pairs that maximizes total merit across all pairings. The total merit for the T-purple/P-white assignment is greater than the alternative, T-white/P-purple. Thus, observers infer that trash maps to purple, despite trash being more strongly associated with white.

The ability to perform assignment inference depends on *semantic discriminability* of the colors, given the concepts in the encoding system. Semantic discriminability can be understood by analogy to perceptual discriminability. Perceptual discriminability concerns how well one can distinguish the appearance of different colors, whereas semantic discriminability concerns how well one can distinguish the meaning of different colors in the context of an encoding system [22, 36]. In

### A. Dissociation of inferred mappings and direct associations

| Concepts | T | P |
|----------|---|---|
| Colors   |   |   |
| TRASH    |   |   |

#### Inferred mapping

- **Stronger associate**

Fig. 2: (A) Example of dissociation between assignments and inferred mappings from [38] (figure adapted from [22]). Bipartite graph represents associations for concepts trash (T) and paper (P) with colors purple and white (thicker edges mean stronger associations). Observers infer trash maps to purple even though trash is more strongly associated with white, which is the optimal global assignment. (B) Color pairs with high vs. low semantic discriminability for concepts watermelon (W) and mango (M) from [36]. ∆S indicates semantic distance.

Assignment inference, semantic discriminability is the degree to which one assignment has greater merit than the alternative assignment(s). For example, Fig. 2B shows color sets that differ in semantic discriminability for the concepts mango (M) and watermelon (W) (data from [36]). The red and orange set (left) has high semantic discriminability because the W-red/M-orange assignment has far greater merit than the alternative. In contrast, the red and green set (right) has low semantic discriminability because the W-red/M-green assignment is only slightly better than the alternative. In their semantic discriminability theory, Mukherjee et al. [22] specified constraints on the ability to design semantically discriminable color palettes for a given set of concepts.

Semantic discriminability can be operationalized through a metric called *semantic distance* (∆S) [22, 36], which uses merit to quantify the degree to which any one assignment is better than alternative assignment(s), while accounting for uncertainty in the system. We reproduce the details for calculating semantic distance defined in [36] in Supplementary Material Section S.2 of the present paper.

#### Simulating assignment inference

To simulate the outcome of assignment inference, it is necessary to (a) determine which assignment is optimal according to an assignment problem [23] and (b) estimate the probability of inferring any one assignment over all alternative assignment(s), which is given by semantic distance. The combination of these two pieces of information indicates which colors observers will map to which concepts in assignment inference, and the probability that they will infer that assignment. This method is effective for predicting how people map colors to concepts for visualizations of categorical information (e.g., recycling bin signage [38], bar charts [22, 36]), although earlier work did not yet refer to “semantic distance” by name [38]. This approach may also extend to inferences about properties of food and beverage products based on coloring in package design [46].

#### Definitions of merit for direct associations

So far, we have treated merit merely as direct association strength. However, there are multiple methods to specify merit for direct associations, with some more effective than others [38]. These different methods reduce to the same outcome in encoding systems with two concepts and two colors, such as those modeled in the present paper. Thus, we will withhold further discussion of metrics for computing merit for direct associations here, and we refer the interested reader to [38] and [22].

#### 2.2 Relational associations

Relational associations are correspondences between relational properties of visual features (e.g., darkness, opacity, spatial arrangement) and relational properties of concepts (e.g., concepts of greater or lesser quantities). A fundamental aspect of relational associations is that they are structure-preserving. Structure preservation arises when structural properties between visual features correspond to structural properties among the concepts to which they are mapped [3, 11, 14, 20, 29, 42, 50].
If particular relations among visual features are salient and certain relations among represented features are salient, then correspondences between these relations can be exploited to constrain the number of potential inferred mappings. For example, people are sensitive to the natural progression from lighter to darker shades and to the natural progression from smaller to larger quantities. Lightness can be mapped to quantities in many ways (see Fig. 3A for four of many possibilities), but only two ways are structure preserving: darker colors map to larger quantities (dark-more assignment) or lighter colors map to larger quantities (light-more assignment). From the perspective of structure preservation, both assignments are equally “good.” Any assignment that scrambles the mapping of lightness values to quantities is not structure-preserving and thereby is less “good.”

Yet, not all structure-preserving assignments are equally good in people’s inferred mappings. People have biases prioritizing one structure-preserving assignment over another [8, 21, 35, 43], discussed below.

**Dark-is-more bias.** The dark-is-more bias is the expectation that darker colors map to larger quantities (“more” of what is being measured) [8, 21, 35, 43]. People have a robust dark-is-more bias when interpreting colormaps without legends [8, 21] and with legends [35, 43]. Studying visualizations without legends, McGranaghan [21] asked participants to interpret maps of U.S. states colored in shades of blue, and found that participants inferred that darker blues mapped to “more.” McGranaghan [21] was purposefully ambiguous about the concept represented in the visualization, stating that the maps represented different amounts of “data” to avoid effects of direct color-concept associations. Studying visualizations with legends, Schloss et al. [35] presented participants with colormaps representing alien animal sightings, with the assumption that people would not have direct associations with these novel concepts. The legend either indicated dark-more encoding (greater animal sightings mapped to darker colors) or light-more encoding (greater animal sightings mapped to lighter colors). Overall, participants were faster at correctly interpreting the visualizations when legends indicated dark–more encoding, compared to light–more encoding, providing further evidence for the dark–is–more bias.

**Opaque-is-more bias.** The opaque–is–more bias is the expectation that regions appearing more opaque represent larger quantities. This bias is only applicable when visualizations appear to vary in opacity [2, 35], such as in value-by-alpha maps [33]. When the opaque–is–more bias is activated, it aligns with the dark–is–more bias on light backgrounds but conflicts with the dark–is–more bias on dark backgrounds. Under such conflicts, the opaque–is–more bias can cancel or even override the dark–is–more bias, leading observers to infer that lighter colors map to larger quantities [2, 35]. When the opaque–is–more bias is non-applicable (i.e., a visualization does not appear to vary in opacity), the dark–is–more bias leads observers to infer that darker colors map to larger quantities on both dark and light backgrounds [2, 35].

**Hotspot–is–more bias.** The hotspot–is–more bias is the expectation that spatial regions that look like hotspots represent larger quantities in data. Hotspots emerge in datasets like fMRI, EEG, and meteorological data, in which extreme values are neighbored by less extreme values in concentric ring-like patterns [40]. Sibrel et al. [43] found that the dark–is–more bias dominated over the hotspot–is–more bias unless the hotspot was highly salient. Still, when colormaps contained hotspots that encoded larger quantities, they were easier to interpret when the hotspot was dark than when it was light (i.e., dark–is–more bias) [43].

3 **Current approach**

Previous work on assignment inference focused on visualizations of categorical information, where merit depends on direct associations [22, 36, 38]. We propose that assignment inference also governs inferred mappings for visualizations of continuous data, where merit may depend on both direct and relational associations. As such, assignment inference would operate over multiple (sometimes competing) sources of merit to determine inferred mappings.

To test this possibility, we asked participants to infer the meanings of colors in colormaps (Fig. 1), and then predicted their responses using simulations of assignment inference. We studied inferred mappings for colormaps without legends, similar to [8, 21]. We assessed the proportion of times participants inferred the darker region mapped to “more,” depending on the domain concept and the color scales used to construct the colormap. In Fig. 1, the domain concept is ocean water, and participants indicated whether there was more ocean water on the left or right of the maps. Colormaps were displayed on a white background and avoided hotspot spatial structure to prevent cases in which the dark–is–more bias conflicted with the opaque–is–more bias [35] and hotspot–is–more bias [43]. In the General Discussion, we discuss extending our approach to handle these additional biases.

Next, we consider how direct and relational associations can serve as sources of merit for visualizations of continuous data, and how multiple sources combine to produce inferred mappings in assignment inference.

### 3.1 Direct associations as a source of merit

Representing merit for direct associations in assignment inference for visualizations of continuous data (Fig. 1) is analogous to representing merit for direct associations for visualizations of categorical information (Fig. 2). In the examples in Fig. 1, merit from direct associations for the colormaps is illustrated in the bipartite graphs under the label “direct associations.” In the bipartite graphs, circles represent the endpoint concepts (more ocean water; +0, and less ocean water; -O) and squares represent the endpoint colors of color scales used to create the colormaps. Edge thickness represents association strength between each endpoint color and concept. From the perspective of merit from direct associations alone, assignment inference simulations for the colormaps in Fig. 1 predict that more ocean water should map to darker blues in the top row and should map to lighter blues in the bottom row.

Although the colormaps represent continuous data (more vs. less ocean water) with a continuous gradation of color, we simplify the assignment problem by focusing on only the endpoint concepts and endpoint colors. As described in Section 2.1, merit for direct associations can be computed in multiple ways, but they simplify to the same outcome when there are two colors and two concepts [38]. By limiting our simulations to the two endpoint colors and two endpoint concepts, we can think about merit for direct color-concept associations simply as association strength. This simplification assumes that colors between the endpoints vary monotonically in association strength with the domain concept (e.g., ocean water in Fig. 1).

### 3.2 Relational associations as a source of merit

To consider how relational associations can be represented as sources of merit in assignment inference for visualizations of continuous data, we first turn to Figs. 3B–C. In these bipartite graphs, edges connect each possible color (shades from white to black) to each possible concept (numeric values from 1 to 4). As indicated in Fig. 3B, only two possible sets of edges are structure-preserving with respect to the natural orderings of quantity and lightness: the set representing dark–more assignment (colored black) and the set representing light–more assignment (colored blue). Edges within each structure-preserving assignment receive the same merit (i.e., edges that are not structure-preserving (colored gray), assuming that each set of structure-preserving edges is bound together (e.g., all blue or all black edges) and never a mix (e.g. some blue and some black edges). Based on structure preservation alone, dark–more and light–more assignments have equal merit, and thus should not be semantically discriminable.

Fig. 3C shows merit from the combination of structure preservation and the dark–is–more bias. The dark–is–more bias places additional merit on structure-preserving edges representing dark–more assignment (thicker edges in Fig. 3C). With greater merit on the dark–more assignment than light–more assignment, these two structure preserving
We addressed this question using colormaps depicting fictitious data to evaluate how weight on direct associations than on the dark-is-more bias (see Exp. 3). Our approach assumed the colors in color scales used to construct data can be simulated using a weighted sum over multiple sources of merit from the dark-is-more bias alone, assignment inference simulations for these colormaps predict that more ocean water should map to darker blues in the top row (consistent with direct associations) and darker browns in the bottom row (conflicting with direct associations).

Like for direct associations, we reduced the problem to model only the endpoint colors and concepts. This simplification ensured that the edges from each potential structure-preserving set (dark-more and light-more) are not mis-matched during simulations of inferred mappings. Our approach assumes the colors in color scales used to construct colormaps vary monotonically in lightness, which was true in the present study (we return to this issue in the General Discussion).

3.3 Combining direct and relational sources of merit

We propose that assignment inference for visualizations of continuous data can be simulated using a weighted sum over multiple sources of merit. With knowledge on how much weight to put on merit from direct associations ($W_d$) and the dark-is-more bias ($W_P$), we can combine these sources of merit (combined merit bipartite graph in Fig. 1) and use established methods for simulating assignment inference [22,36,38] with the specifications of their monitors. This experiment took approximately 30 min. and participants were compensated $3.63. The mean color for each of the UW-71 colors [22] shown in Fig. 4B (see Table S.3 in Supplementary Material for CIELAB coordinates). The UW-71 colors include 58 colors uniformly sampled from CIELAB space (UW-58 from [31, 36]), plus 13 additional colors sampled at a higher lightness plane to incorporate more saturated yellows and greens [22].

Our target sample size was $n = 30$ and we collected data from 35 Amazon mTurk workers given we expected several participants would be excluded for failing the attention check, described below (35 total, 3 excluded). The final sample was $n = 32$ (mean age = 40 years old; 11 women, 21 men; gender assessed using free-response here and in all subsequent experiments). All participants indicated normal color vision for independent effects of each factor.

Before beginning the task, participants were shown the domain concept words and all 71 colors (Fig. 4B). They were asked to identify which color they associated most with the given color by moving a slider along a scale ranging from “not at all” (−200) to “very much” (200), and clicking “continue” to begin the next trial. Each concept was rated for each of the UW-71 colors [22] shown in Fig. 4B (see Table S.3 in Supplementary Material for CIELAB coordinates). The UW-71 colors include 58 colors uniformly sampled from CIELAB space (UW-58 from [31, 36]), plus 13 additional colors sampled at a higher lightness plane to incorporate more saturated yellows and greens [22].

Our target sample size was $n = 30$ and we collected data from 35 Amazon mTurk workers given we expected several participants would be excluded for failing the attention check, described below (35 total, 3 excluded). The final sample was $n = 32$ (mean age = 40 years old; 11 women, 21 men; gender assessed using free-response here and in all subsequent experiments). All participants indicated normal color vision for independent effects of each factor.

Before beginning the task, participants were shown the domain concept words and all 71 colors (Fig. 4B). They were asked to identify which color they associated most with the given color by moving a slider along a scale ranging from “not at all” (−200) to “very much” (200), and clicking “continue” to begin the next trial. Each concept was rated for each of the UW-71 colors [22] shown in Fig. 4B (see Table S.3 in Supplementary Material for CIELAB coordinates). The UW-71 colors include 58 colors uniformly sampled from CIELAB space (UW-58 from [31, 36]), plus 13 additional colors sampled at a higher lightness plane to incorporate more saturated yellows and greens [22].
4.1.2 Generating colormaps

To generate the colormaps for this experiment, we (1) specified eight pairs of endpoint colors, (2) interpolated between the eight endpoint colors to create color scales, and then (3) applied the color scales to 10 underlying datasets to create colormap data visualizations (Fig. 5).

1. We selected endpoint colors such that one color was lighter (L) and the other was darker (D). For four endpoint pairs, association difference was high—sunshine was far more associated with the light than the dark endpoint, and shade was far more associated with the dark than the light endpoint. For four other pairs, association difference was low, which occurred when both colors were either weakly or moderately associated with the domain concept. Within each level of association difference, two endpoint pairs had a lightness difference of $L^* = 38$, and two had a difference of $L^* = 50$. We tested multiple color pairs for each condition to ensure our results were not specific to any one color pair. We checked if the colors interpolated between the endpoints varied (approximately) monotonically in direct association strength for both sunshine and shade (i.e., the domain concept was not more associated with the intermediate color than with either endpoint). See Supplementary Material Section S.4 and Fig. S.5 for details.

2. Using these endpoint colors, we created eight color scales by linearly interpolating eight steps between the light and dark endpoints (interpolation computed in CIELAB space). The resulting color scales had 10 steps, as in the stimuli from [35].

3. Finally, we applied each of the eight color scales to 10 underlying datasets, producing 80 colormap data visualizations. The underlying datasets produced colormaps appearing as an 8 × 8 grid, where one side was biased to be lighter and the other side was biased to be darker. Within the 10 underlying datasets, half produced colormaps in which the left side was darker than the right side (as in Fig. 5), and the other half produced colormaps in which the right side was darker (as in Fig. 1). A full set of 10 colormaps from one color scale are shown in Fig. S.8 in the Supplementary Material.

The underlying datasets we used were previously used to generate colormaps in [35]. The data ranged from 0-1, with values sampled from eight discrete points along an arctangent curve with added noise. The eight points corresponded to the eight columns of the colormaps. The samples at each point were used to assign values to the rows within each column of the colormap (see Supplementary Material Section S.1 for further details). One endpoint of the color scale was assigned a data value of 0 and the other endpoint a data value of 1, such that the color scales corresponded to the full range of the underlying data. Given that the data were evenly sampled along the arctangent curve, the data represented in the colormaps evenly span the full data range. This method of generating stimuli mitigates concerns about the dynamic range of data variability being hidden in the data visualization [10,52].

4.1.3 Assessing inferred mappings for colormaps

In the colormap task, participants were presented with colormaps along with a domain concept (sunshine or shade). They were told that the colormaps represented amounts of sunshine (or shade) from various counties in a state. In some counties, there was more sunshine (shade) on the left side of the county; in other counties, there was more sunshine (shade) on the right side. Their task was to indicate whether there was more sunshine (shade) on the left/right of the map by pressing the left/right key on their keyboard.

Domain concept and color scale varied between-subjects, and participants were randomly assigned to one of 16 groups (8 color scales × 2 domain concepts). Each participant judged all 10 colormaps for their assigned domain concept and color scale, one at a time in a random order. Trials were separated by a 500-ms inter-trial interval. The colormaps (approx. 4cm × 4cm) appeared on a white square (approx. 9cm × 9cm) in the center of a medium gray screen (size estimates using a 15.6in. 1920 × 1080 pixel monitor). Below each half of the colormap was a horizontal line labeled “Left?”/“Right?” (Fig. 1).

Our target sample size was $n = 192$, $n = 12$ per group (sample size based on a power analysis reported in Supplementary Material Section S.6). The final sample was 187 mTurk workers (mean age = 38.9 years old, 105 women, 82 men), after excluding $n = 3$ for atypical color vision and $n = 2$ for not completing the experiment. The groups ranged from 9 – 13 participants due to how the experiment code automated assignments to conditions while managing exclusions. The experiment took approx. 5 min. and participants were compensated with $0.60.

4.2 Results and Discussion

Fig. 6A shows potential patterns of inferred mappings if there was an effect only of direct associations, only of the dark-is-more bias, or both (right). (B) Results of Exp. 1, showing the mean proportion of times the darker side was selected for maps about sunshine (circles) and shade (triangles) as a function of signed association difference. Mark colors represent endpoint colors in the color scales and error bars represent standard errors of the means.
the domain concept. Dark-is-more only predicts participants would always choose the darker side, regardless of association difference. If both have an effect, then participants would choose the darker side when the two biases are consistent (positive association differences) but less likely to choose the darker side as the lighter side becomes more associated with the domain concept (negative association differences).

Fig. 6B shows the mean proportion of times the darker side was chosen, averaged over the 10 repetitions within each participant, and then averaged over participants. The pattern of responses resembles the predicted pattern if both the dark-is-more bias and direct associations influenced inferred mappings. Participants almost always chose the darker side for shade (association differences greater than zero), and their likelihood of choosing the darker side decreased as the lighter side became more associated with sunshine.

To test for independent effects of each potential source of merit, we used a mixed-effect logistic regression model. Although we plot the data in terms of the proportion of times participants chose the darker side (Fig. 6B), this way of coding the data poses a problem for including the dark-is-more bias as a predictor in a regression model, given that there is no variability in the predictor (it predicts a response of ‘1’ on every trial). Thus, we conducted a model to predict whether participants chose the left side on each trial (1 = left, 0 = right), from a predictor coding whether the left side was darker (1 = left darker, −1 = right darker), and a predictor coding which side was more associated, and by how much (scaled to range from −1 to 1; x-axis values in Fig. 6B). Conducting models with respect to the left side is a standard approach in psychophysics research, and is valid as long as the stimuli are left/right balanced, as in the present stimulus set (see Section 4.1.2).

Participants were more likely to select the left side if it was more strongly associated with the domain concept than the right side ($B = 4.51, SE = 0.21; z = 21.22, p < .001$) and if it was dark than light ($B = 1.33, SE = 0.09; z = 15.46, p < .001$) (dark-is-more bias). Thus, both direct and relational associations influenced inferred mappings. See Section 5.7 in Supplementary Material for an additional analysis that includes concept as a factor in the model.

Summary. Exp. 1 showed that direct associations and the dark-is-more bias contribute independently to people’s inferred mappings. When these two factors conflict (the domain concept is more associated with the light endpoint than the dark endpoint) and the direct association difference is large, direct associations override the dark-is-more bias.

5 Experiment 2

Given evidence that direct associations can override the dark-is-more bias when they conflict and direct associations are strong, we conducted Exp. 2 to test how much association difference was needed for direct associations to fully override the dark-is-more bias. The results led us to study effects of semantic distance for predicting inferred mappings.

5.1 Methods

The methods were the same as Exp. 1, except for two changes. First, we only tested sunshine as the domain concept in order to focus on cases where direct associations and the dark-is-more bias conflict. Second, we included eight new color scales of intermediate association difference, in addition to the original eight from Exp. 1 (16 color scales) (Fig. 7).

For the new color scales, we selected endpoint color pairs using the association data from Exp. 1 with the same selection criteria (Supplementary Material Section S.4). Two of the new color scales did not meet our statistical criterion for monotonocity due to a coding error treating hue angle as radians instead of degrees (Supplementary Material Fig. S.6B), and visual inspection showed that intermediate colors were more associated with sunshine than the endpoints. Thus, we excluded data from these two color scales from analysis.

We focused on association difference and allowed association strength to vary (e.g., the same value of association difference could be achieved if sunshine was moderately associated with the light endpoint and weakly associated with the dark endpoint, or strongly associated with the light endpoint and moderately associated with the dark endpoint). As in Exp. 1, each participant judged 10 colormaps constructed from one of the 16 color scales (between-subjects).

Our target sample size was $n = 640$ ($n = 40$ per condition) based on power analysis (see Supplementary Material Section S.6). We collected data until each condition reached at least 40 participants after excluding those with atypical color vision (696 Amazon mTurk workers in total, 41 excluded). We assessed color vision using the two self-report questions in Exp. 1, plus responses to six digital Ishihara plates. Participants were excluded if they answered yes to either question and/or answered incorrectly for more than one of the six plates. The final sample included 655 participants (mean age = 38.6 years old; 294 women, 358 men, 2 non-binary, 1 no report). The groups ranged from 40 – 44 participants due to how the experiment code automated assignments to conditions while managing exclusions. The experiment took approximately 5 min. and participants were compensated $0.60.

5.2 Results and Discussion

Fig. 8A shows the mean proportion of times participants selected the darker side for colormaps generated from each color scale (averaged over the 10 colormaps judged by each participant, and then averaged over participants). As in Exp. 1, direct associations were more likely to override the dark-is-more bias as association difference increased ($r(12) = .85, p < .001$).

Our target sample size was $n = 640$ ($n = 40$ per condition) based on power analysis (see Supplementary Material Section S.6). We collected data until each condition reached at least 40 participants after excluding those with atypical color vision (696 Amazon mTurk workers in total, 41 excluded). We assessed color vision using the two self-report questions in Exp. 1, plus responses to six digital Ishihara plates. Participants were excluded if they answered yes to either question and/or answered incorrectly for more than one of the six plates. The final sample included 655 participants (mean age = 38.6 years old; 294 women, 358 men, 2 non-binary, 1 no report). The groups ranged from 40 – 44 participants due to how the experiment code automated assignments to conditions while managing exclusions. The experiment took approximately 5 min. and participants were compensated $0.60.

One possibility is that participants approached this task using assignment inference, comparing each possible assignment (dark-more or light-more), and inferring the assignment with greater merit. Once the merit of one assignment is sufficiently greater than the alternative, the colors reach maximal semantic discriminability. Then, further increasing association difference has no further effect on assignment inference. This limit may have been reached at an association difference of around −55. If so, then the plateauing function in Fig. 8A may become linear when we replace the x-axis (signed association difference) with simulations of assignment inference (Section 2.1).

To simulate assignment inference for each color scale, we first calculated semantic distance for each pair of endpoint colors (using equation 1 defined in [36] and reproduced in Supplementary Material Section S.2 of this paper). We then determined the optimal assignment (i.e., which assignment had greater merit), and coded the outcome as dark-more = +1 and light-more = −1. Last, we multiplied this coding by semantic distance to compute signed semantic distance, which gave positive values to the probability of inferring dark-more assignments and negative values to the probability of inferring light-more assignments.

Computing signed semantic distance required specifying merit based on direct associations between each endpoint color and each endpoint concept (“no sunshine” vs. “a lot of sunshine”), as in Fig. 1. From Exp. 1, we had association data for domain concept “sunshine,” but not the endpoint concepts. Thus, we collected data from additional participants ($n = 31$), who rated the association strength between each endpoint color and endpoint concepts “no sunshine” and “a lot of sunshine” (see Supplementary Material Section S.5). We used the mean ratings as merit to compute signed semantic distance.

As shown in Fig. 8B, inferred mappings were predicted by simulations of assignment inference: signed semantic distance was strongly...
We first collected direct association ratings for the domain concepts. We shortened the attention check (70 Amazon mTurk workers total. We shortened the attention check in Material Section S.3). Our final sample was a priori color vision (n = 15) and who failed the attention check (n = 17); 70 Amazon mTurk workers total. We shortened the attention check block but used the same a priori exclusion criterion (see Supplementary Material Section S.3). Our final sample was n = 38 (mean age = 42.8 years old, 18 women, 20 men). The experiment took approximately 60 min. and participants were compensated $7.25.

6.1.2 Generating colormaps and computing merit

Based on the direct association data, we created colormaps for three concepts: wild fire, ocean water, and glacial ice. These domain concepts enabled spanning the full range of signed semantic distances within each concept. We generated colormaps using the methods in Exp. 1 (Fig. 5) and Supplementary Material Section S.4. For each domain concept, we chose 21 pairs of endpoint colors that spanned the full range of association differences from strongly negative (light color was more associated with the domain concept) to strongly positive (dark color was more associated with the domain concept). Fig. 9 shows the resulting 21 color scales for each domain concept. All color scales satisfied the criterion for monotonicity. As in Exps. 1 and 2, each color scale was applied to 10 underlying datasets to produce 10 colormaps per color scale, with darker side left/right balanced (Fig. S.8).

After selecting the color pairs, we collected additional data to estimate merit for each endpoint color paired with each endpoint concept, with respect to direct associations and the dark-is-more bias (Fig. 1).

Merit for direct associations. A new set of 30 participants rated the association strength between each endpoint of each domain concept (e.g., “a lot of ocean water,” “no ocean water”) and each corresponding endpoint color (details in Supplementary Material Section S.5). As in Exp. 2, we used these associations to estimate merit derived from direct associations for each color-concept endpoint pairing (Fig. 1).

Merit for the dark-is-more bias. So far, we have discussed the dark-is-more bias as binary—dark-more assignments have greater merit than light-more assignments. However, the dark-is-more bias can also be considered continuous—the degree to which dark-more assignments have greater merit depends on the degree to which one endpoint appears clearly darker than the other endpoint. One might consider quantifying merit of the dark-is-more bias using lightness (L*) difference between the two endpoint colors of the color scale that varied monotonically in lightness. However, we reasoned that the dark-is-more bias would be activated if one side appeared clearly darker than the other, and adding additional lightness difference may not increase activation of the bias.

Thus, we used an empirical approach to quantify merit for the dark-is-more bias. For each endpoint color pair, volunteers with expertise in color perception (n = 4) rated the degree to which one color was clearly darker than the other color (referred to as darkness difference ratings). They judged each pair twice (left/right balanced), and made their ratings on continuous slider scale from “left color is clearly darker” to “right color is clearly darker." The middle was labeled “equal darkness” (see Supplementary Material Section S.5 for details). For each color scale, we coded dark-more edges to have merit = 1 and light-more edges to have merit = 0, and then multiplied these values by the darkness difference ratings. As a result, differences in total merit of dark-more vs. light-more assignments scaled with the degree to which it was obvious that the dark endpoint appeared darker than the light endpoint.
6.1.3 Colormap interpretation task

This task was the same as in Exps. 1 and 2, except the domain concepts were wild fire, ocean water, and glacial ice and there were 21 color scales per domain concept (3 domain concepts \( \times 21 \text{color scales} = 63 \text{groups of participants} \)). Participant judged 10 colormaps for their assigned domain concept and color scale. They were told that the colormaps represented data about [domain concept] in different counties. Their task was to indicate whether there was more [domain concept] on the left/right side of the county (Fig. 1).

The target sample size was \( n = 1260 \) (\( n = 20 \) per condition) based on a power analysis (see Supplementary Material Section S.6). We collected data in batches until each condition had at least \( n = 20 \) after excluding those with atypical color vision as assessed in Exp. 2 (1391 Amazon mTurk workers total, 107 excluded). The final sample was \( n = 1284 \) (mean age = 40.3 years old, 1 no reported age; 672 women, 598 men, 9 non-binary, 5 no reported gender). The groups ranged from 20 – 22 participants due to how the experiment code automated assignments to conditions while managing exclusions. The experiment took approximately 5 min. and participants were compensated $0.60.

6.2 Results and Discussion

In the following analyses, we determined the optimal relative weighting on direct and relational associations, and then assessed whether assignment inference simulations using the optimal weighting predicted inferred mappings better than simulations using each source of merit alone. We split participants into a training set to determine the optimal weighting, and a test set to compare the optimal weighting with each source of merit alone. Each set had 10 – 12 participants per color scale for each domain concept. We simulated assignment inference with varying relative weight on each source of merit as follows:

1) Computing combined merit. First, we specified merit of each color-concept pairing within each source of merit (Section 6.1.2). Then, we calculated combined merit by computing the weighted sum over bipartite graphs for each source of merit (Fig. 1). We used each combination of weights on direct associations (\( W_d \)) and the dark-is-more bias (\( W_p \)) in increments of .05, such that their sum was 1. Each weight was a multiplicative factor on each edge of the respective bipartite graphs. For instance, a weight pairing of (.1, 0) placed all the weight on direct associations, (.0, 1) placed all the weight on the dark-is-more bias, and (.5, .5) placed equal weight on both sources of merit.

2) Computing signed semantic distance. We computed signed semantic distance over combined merit for each weight pairing, following the procedure in Exp. 2. First, we computed semantic distance between the endpoint colors for each domain concept. Next, we determined the optimal assignment (which assignment had greater overall merit), coded as +1 for dark-more and −1 for light-more. Last, we multiplied this coding by semantic distance to obtain signed semantic distance.

To determine the optimal weighting, we used mean squared error (MSE) to compare assignment inference simulations with human judgments. For each of the 21 color scales for each of the three domain concepts, we computed MSE between mean responses (scaled to range from 0 to 1) and signed semantic distance with the optimal weight pair. Mark colors indicate color scale 

\[ r = 0.83 \]

Fig. 11 shows the relation between participant responses and simulations of assignment inference using the optimal weight pairings for each color scale. Points would fall along the diagonal line if the simulations perfectly predicted inferred mappings. Signed semantic distance was significantly correlated with inferred mappings for all three domain concepts, but to varying degrees: strong correlation for wild fire (\( r = .83, p < .001 \)), moderately strong for ocean water (\( r = .72, p < .001 \)), and moderate for glacial ice (\( r = .55, p = .01 \)). Preliminary exploration suggests this weaker relation for glacial ice might be due to some colormaps appearing to vary in opacity, activating the opaque-is-more bias. The opaque-is-more bias aligns with the dark-is-more bias on light backgrounds (as used here), and the two relational associations may have combined to jointly override effects of direct associations. Our study was not designed to test the opaque-is-more bias, so future work is needed to study these effects more directly.\(^6\)

Supplementary Material Section S.7 includes an additional analysis showing independent effects of direct and relational associations on

\(^6\)This weaker correlation for glacial ice is not likely due to concept glacial ice, but rather the colormaps appearing to vary in opacity happened to be in the glacial ice condition. The applicability of the opaque-is-more bias depends on the combination of background and colors of the color scale [35], and would be applicable for these colormaps if they represented any other domain concept.
inferred mappings (as in Exp. 1). It also includes plots of inferred mappings as a function of association difference and semantic distance from direct associations (Fig. S.9), analogous to Fig. 8 in Exp. 2. The strong plate for sunshine colormaps in Exp. 2 was less apparent in Exp. 3, and we consider possible explanations in Section S.7.2.

Summary. Exp. 3 showed that inferred mappings for colormaps were well-predicted using a simulation of assignment inference with combined merit. The optimal combined weighting resulted in predictions with less error than predictions simulated with weight on direct associations or the dark-is-more bias alone.

7 GENERAL DISCUSSION

A central problem in visual communication is understanding how people infer meaning from visual features. By anticipating people’s expectations about how visual features should map onto concepts, designers can create visualizations that align with those expectations, thereby facilitating communication [14, 18, 22, 26, 35, 36, 38, 43, 50, 51].

We approached this problem by bridging work on inferred mappings for visualizations of categorical information [18, 22, 36, 38, 41] and visualizations of continuous data [8, 21, 35, 43] to understand both within a single framework of assignment inference. Doing so required broadening the notion of merit in assignment inference to include not only direct associations as in [22, 36, 38], but also relational associations (e.g., dark-is-more bias). Exp. 1 showed that direct and relational associations contribute independently to inferred mappings for colormaps. Exp. 2 showed that inferred mappings for colormaps were predicted by simulations of assignment inference (signed semantic distance) using merit from direct associations. Exp. 3 showed that simulating assignment inference using a weighted sum over merit from direct and relational associations was better at predicting inferred mappings than simulations using each source of merit alone.

This study is an initial step towards comprehensively modeling the effects of multiple sources of merit in assignment inference. Here, we began with direct associations and one type of relational association, the dark-is-more bias. In future work, we will extend our approach to include additional sources of merit, including the opaque-is-more bias and hotspot-is-more bias. To quantify merit for the opaque-is-more bias, it will be necessary to estimate the degree to which colors in the colormap appear to vary in opacity depending on the background color (see [35]), and ensure that this source of merit falls out of the equation when colormaps do not appear to vary in opacity. To quantify merit for the hotspot-is-more bias, it will be necessary to quantify the degree to which hotspots are salient in the colormap, and again ensure that this source of merit falls out of the equation when colormaps do not appear to have hotspots. Our approach for estimating inferred mappings not only has potential to accommodate known sources of merit, but can also scale as additional sources of merit are discovered.

We also expect our approach to extend to abstract concepts. Evidence suggests that sets of abstract concepts previously considered “non-colorable” (e.g., sleeping, driving, safety, comfort) can be meaningfully encoded using color as long as their association distributions are sufficiently different from one another (semantic discriminability theory [22]). In the present framework, as long as the colors in the color scale vary in association strength with the domain concept, then merit from direct associations will influence combined merit with the dark-is-more bias. If the associations do not vary in association strength (low semantic distance), then merit from direct associations will have little effect on combined merit, and the dark-is-more bias should dominate inferred mappings. These patterns should hold regardless of whether the concepts are abstract/concrete. If a concept has no systematic color-concept associations, regardless of whether it is abstract/concrete, then it will not be possible to create a color scale with large direct association difference, so the dark-is-more bias (and any other sources of merit) would dominate inferred mappings.

Overall, our findings can be translated to incorporate color semantics into tools that generate colors for information visualizations (e.g., Colorgorical [12], Color Crafter [45], and CCC-Tool [25]). These tools already allow designers to balance different factors, such as perceptual discriminability and aesthetics. With a comprehensive model of assignment inference combining multiple sources of merit, it will be possible to incorporate semantic discriminability into algorithms that optimize color selection for visualization design.

Limitations. This study has limitations for future work to address. Linearly interpolated color scales. We used color scales that were linearly interpolated between two endpoints in CIELAB space, which supported the goals of this study. Interpolated color scales allowed us to compare merit of dark-more vs. light-more assignments using direct color-concept associations from only the endpoint colors. Using only the endpoints was possible because the intermediate colors varied approximately monotonically in association strength between the end-point colors (see Supplementary Material Section S.4). Monotonicity would be violated if the domain concept was more/less associated with intermediate colors of a color scale than the endpoints (e.g., using a color scale for sunshine that interpolated between a red and yellowish-green, resulting in more strongly associated yellows in the middle).

Monotonicity would also likely be violated in industry standard color scales that spiral through color space [5, 16, 45]. Yet, Smart et al. [45] showed that such color scales that spiral produce colormaps that are more interpretable and aesthetically preferable than linear colormaps like the ones in the present study. Indeed, many criteria determine whether color scales (also referred to as ramps) are effective for visualizing continuous data [6, 25, 32, 34, 44, 45, 53] and our color scales were not designed to meet those criteria. Thus, the color scales in the present study are not meant to be used for visualizations of real data. To apply our modeling approach to more complex color scales, it will be necessary to quantify merit for color-concept pairings sampled in multiple steps between the two endpoints, and use a method for computing the optimal assignment that accounts for many colors and many concepts.

Sequential color scales. The present work, and most previous work on inferred mapping for colormaps [21, 35, 43], has focused on sequential color scales, where encoded data ranged from small to large. Questions remain concerning how this work extends to diverging scales, where encoded data has a neutral point. The dark-is-more and opaque-is-more biases imply that more extreme data (furthest from the neutral point) should map to darker, more opaque regions, respectively. Future work is needed to test these hypotheses, and to investigate people’s expectations about which colors represent data values above/below the neutral point. Future work is also needed to determine whether the relative weightings on sources of merit established in Exp. 3 for sequential color scales generalize to diverging color scales.

Task type. Our study and much of the previous studies on inferred mappings for colormaps [21, 35, 43] used tasks that asked participants to interpret where “more” was represented in a colormap. However, people use colormap data visualizations for a wide variety of other tasks, such as those studied by Padilla et al. [28]: finding a specific value of a concept, comparing values across regions, and averaging values across regions. Future research is needed to test whether the present framework modeling combined merit to estimate inferred mappings predicts performance in these other kinds of tasks.

Conclusion. This work builds a new bridge for understanding how direct and relational associations combine to influence inferences about the meanings of colors in visualizations. We have laid the groundwork to develop a more comprehensive model of assignment inference that accounts for additional sources of merit that we know of, and can scale to accommodate new sources of merit as they are discovered. Our findings can be translated directly to design visualizations that align with people’s expectations about the meanings of colors, thereby making visualizations that are easy to interpret.

ACKNOWLEDGMENTS

We thank Kushin Mukherjee, Chris Thorstenson, Anna Bartel, Clementine Zinnicki, Madeline Parker, and Audrey Wang for helpful discussions on this work. This project was supported by the UW–Madison Office of the Vice Chancellor for Research and Graduate Education, Wisconsin Alumni Research Foundation, NSF award BCS-1945303 to KBS, and NSF award #2050782.
[52] Q. Zeng, Y. Zhao, Y. Wang, J. Zhang, Y. Cao, C. Tu, I. Viola, and Y. Wang. Data-driven colormap adjustment for exploring spatial variations in scalar fields. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2021.

[53] L. Zhou and C. D. Hansen. A survey of colormaps in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(8):2051–2069, 2016.