Unifying Effects of Direct and Relational Associations for Visual Communication

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

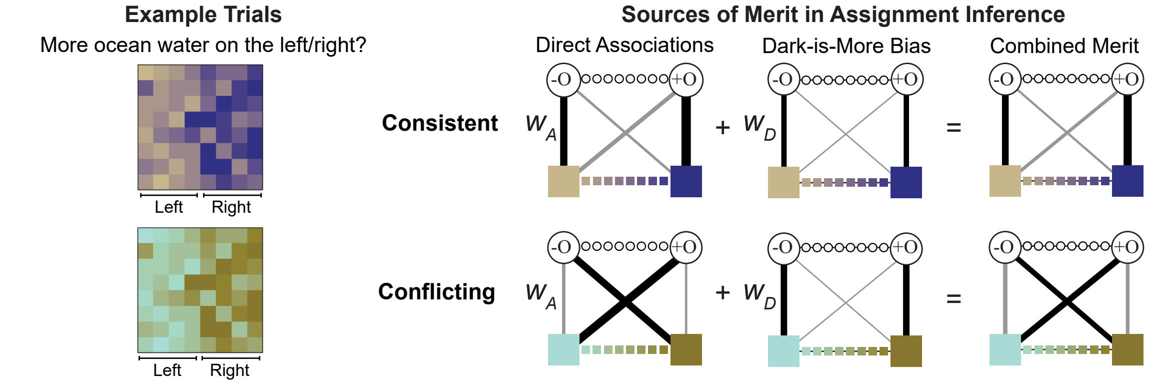

Fig. 1: In this study, participants inferred which region of colormaps (left/right) represented more of a domain concept (e.g., ocean water). Inferences can be predicted by simulating assignment inference using a weighted combination of multiple (sometimes competing) sources of "merit": direct associations and relational associations (dark-is-more bias).

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., mangos are associated with yellows) 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 concepts represented in the chart. Now, imagine instead interpreting a colormap data visualization[1] and you need to infer which colors map to which quantities represented in the colormap.

- *Melissa A. Schoenlein, Psychology and Wisconsin Institute for Discovery, University of Wisconsin–Madison. Email: schoenlein@wisc.edu.*
- *Johnny Campos, Cognitive Science, University of California, Merced. Email: jcampos54@ucmerced.edu.*
- *Kevin J. Lande, Philosophy and Centre for Vision Research, York University. Email: lande@yorku.ca*
- *Laurent Lessard, Mechanical and Industrial Engineering, Northeastern University. Email: l.lessard@northeastern.edu*
- *Karen B. Schloss, Psychology and Wisconsin Institute for Discovery, University of Wisconsin–Madison. Email: kschloss@wisc.edu.*

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

[1] Literature 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.

Manuscript received 31 March 2022; revised 1 July 2022; accepted 8 August 2022. Date of publication 29 September 2022; date of current version 2 December 2022. This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2022.3209443, provided by the authors. Digital Object Identifier no. 10.1109/TVCG.2022.3209443

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 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.

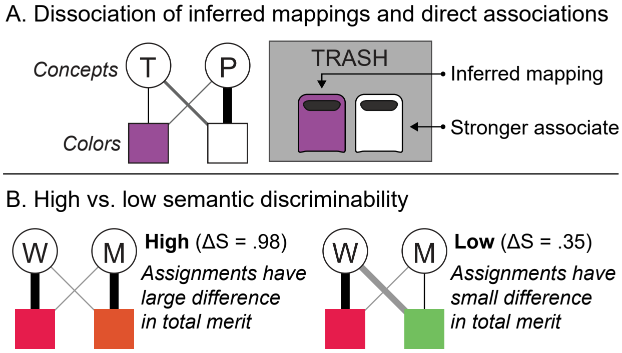
A. Dissociation of inferred mappings and direct associations
Concepts
T
P
Colors
TRASH
Inferred mapping
Stronger associate
B. High vs. low semantic discriminability
W
M
High (ΔS = .98)
Assignments have large difference in total merit
W
M
Low (ΔS = .35)
Assignments have small difference in total merit

Fig. 2: (A) Example of dissociation between associations 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]. ΔS indicates semantic distance.

Semantic discriminability can be operationalized through a metric called ***semantic distance (ΔS)*** [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 peoples' 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][2] 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].[3] 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; +O, 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 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 more merit than 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

[2] Although legends are a central part of colormap visualization grammar, Christen et al. [7] found that journal articles often leave out legends. Thus, studying colormaps without legends is relevant for real-world visualizations, while also providing a direct window into people's inferred mappings.

[3] This method requires fewer within-subject trials per condition than methods assessing inferred mappings for colormaps with legends, which require counterbalancing legend conditions (see [35, 43]).

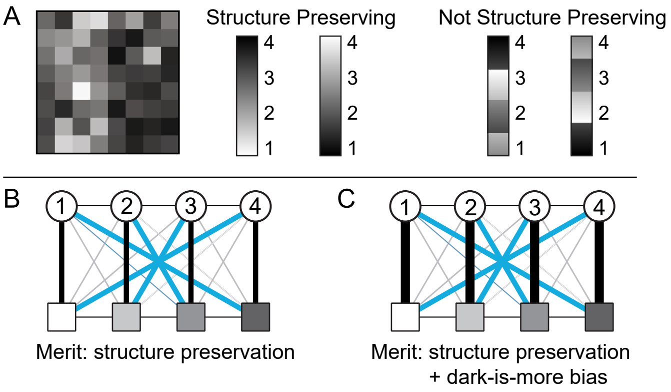
A
Structure Preserving
Not Structure Preserving
4
3
2
1
B
C
1
2
3
4
Merit: structure preservation
Merit: structure preservation + dark-is-more bias

Fig. 3: Illustrations of structure preservation. (A) A colormap assigning lightness (light to dark) to quantity (1-4), with legends specifying structure-preserving assignments (natural progressions of lightness correspond to a natural ordering of quantities) vs. not structure-preserving assignments (assignment of lightness to quantity is scrambled). Bipartite graphs can code merit in terms of (B) structure preservation and (C) structure preservation plus the dark-is-more bias.

assignments should be semantically discriminable. Given that treating the dark-is-more bias as a source of merit also implies structure preservation, here forward, we focus on the dark-is-more bias.

Fig. 1 shows the dark-is-more bias represented as a source of merit for the example colormaps about ocean water. From the perspective 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_A) and the dark-is-more bias (W_D), 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] to predict inferred mappings. In the top row, these sources of merit are consistent, and simulating assignment inference over combined merit predicts observers will infer that darker colors map to larger quantities. In the bottom row, these two sources of merit are conflicting. Depending on the relative weight given to each source, they might cancel each other out, or one might dominate over the other. The weights used in Fig. 1 are based on the results of Exp. 3, with greater weight on direct associations than on the dark-is-more bias (see Exp. 3 for details).

In this study we asked whether direct and relational associations independently contribute to merit in assignment inference for colormap data visualizations, and if so, what is their relative contribution? Answering these questions enabled us to create a model that predicts people's inferred mappings, which can be used to help design colormaps that facilitate visual communication.

4 Experiment 1

Experiment 1 investigated whether both direct and relational color-concept associations contribute to inferred mappings for colormaps. We addressed this question using colormaps depicting fictitious data about two domain concepts, shade and sunshine. We chose these concepts because the dark-is-more bias and direct associations would be consistent for shade and conflicting for sunshine, allowing us to test for independent effects of each factor.

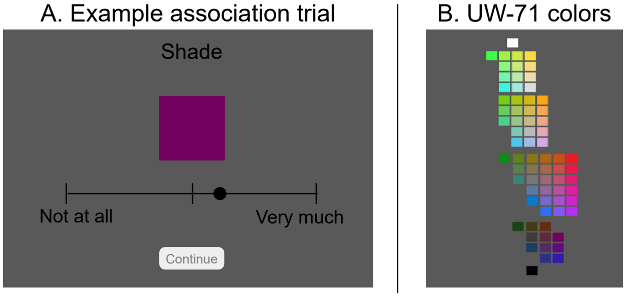

Fig. 4: (A) Example association rating trial. The slider indicates a slight association for the given purple color with the concept shade. (B) The UW-71 colors as seen during the association task instructions.

4.1 Methods

We began by collecting direct color-concept association data for the domain concepts shade and sunshine. We then used these data to generate colormap stimuli to assess inferred mappings. Data, code, and color coordinates for all experiments in this paper can be found at https://github.com/SchlossVRL/assign-infer-colormaps.

4.1.1 Measuring direct color-concept associations

In the color-concept association task, participants were presented with a concept word at the top of the screen (sunshine or shade) and a colored square centered below (Fig. 4A). They rated how much they associated the given concept 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 when asked if they had difficulty distinguishing between colors relative to the average person and if they considered themselves colorblind. All participants of this and all subsequent experiments gave informed consent and the UW–Madison IRB approved the protocol.

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 and least with each concept to anchor the endpoints of the scale [30].The experiment was blocked by concept, with shade and sunshine presented in a random order within the first two blocks. The 71 colors appeared in a random order within each block. Given our plan to use association ratings from this task to generate stimuli for the colormaps task, we sought to include associations only from participants who made careful judgments. Thus, we included a third, attention check block for all participants and set an *a priori* exclusion criterion (see Section S.3 in the Supplementary Material).

The displays of this and all subsequent experiments were created using jsPsych [9]. All participants completed the experiments on their own devices so the color coordinates were calculated using standard assumptions about RGB displays. Thus, as is typical in color experiments in visualization, which aim to be robust to variations across displays [12,22,47,48], the precise colors each participant saw varied with the specifications of their monitors. This experiment took approximately 30 min. and participants were compensated $3.63. The mean color-concept associations for sunshine and shade are shown in Fig. S.4 of the Supplementary Material.

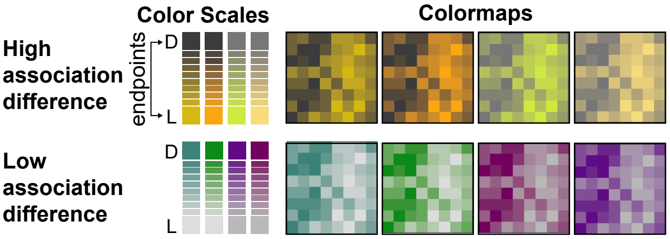

Fig. 5: Color scales and corresponding colormaps in Exp. 1.

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 colorscales 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.[4] 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 colors than with either endpoint)[5]. 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 colormaps 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.

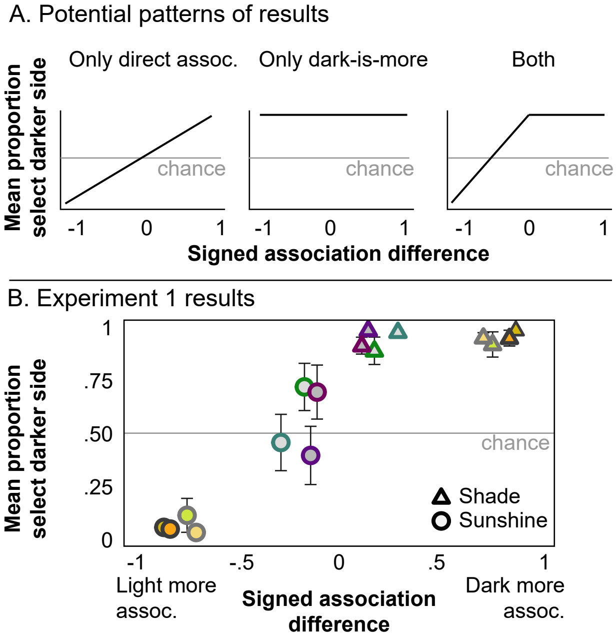

Fig. 6: (A) Predicted patterns of results in Exp. 1 if inferred mappings are influenced by only the dark-is more bias (left), only direct associations (center), or both (right). (B) Results of Exp. 1, showing the mean proportion of times the darker side was selected for maps about sunshine (circles) 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.

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. The y-axis represents the proportion of times the darker side would be chosen over all trials, as a function of signed association difference. Positive/negative association differences indicate the darker/lighter color is more associated with the domain concept, respectively. Direct associations only predicts the probability of choosing the darker side would increase monotonically as the darker side becomes more associated with

[4] Overall, mean association ratings increased with lightness for sunshine (CIELAB L*) ($r(69) = .71, p < .001$) and decreased with L* for shade ($r(69) = -.79, p < .001$), but some light colors were moderately associated with sunshine, and some dark colors were moderately associated with shade. These properties enabled us to generate colormaps that varied in association difference.

[5] Two color scales for shade did not meet our statistical criterion, due to a coding error treating hue angle as radians instead of degrees. However, our statistical criterion is a heuristic, and visual inspection suggested that the intermediate colors still varied monotonically between the endpoints (Supplementary Material Fig. S.6A), so we kept data for these color scales in the analysis.

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 S.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 monotonicity 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).

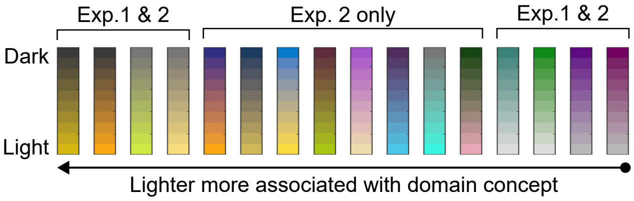

Fig. 7: Color scales used in Exp. 2. From right to left, sunshine is increasingly more associated with the light endpoint of the color scale. The four leftmost and four rightmost color scales were used in Exp. 1.

Our target sample size was $n = 640$ ($n = 40$ per condition) based on a 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 ($r(12) = .85, p < .001$). But, once association difference reached about $-.55$, participants almost always inferred that the lighter side of the colormaps mapped to more sunshine. Direct associations fully overrode the dark-is-more bias, so further increasing association difference could not further influence inferred mappings (floor effect). This observation led us to ask, why would inferred mappings level off at around $-.55$?

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

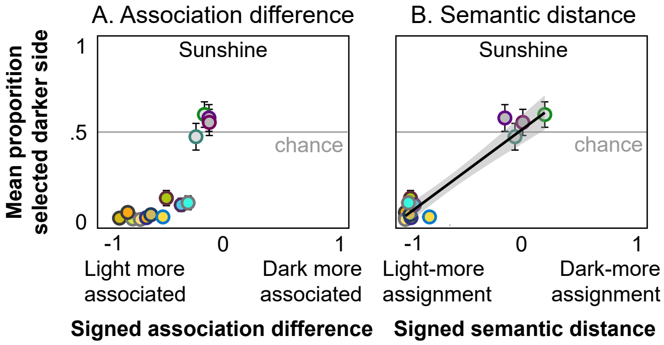

Fig. 8: Mean proportion of times the darker side was selected for each color scale, plotted as a function of (A) signed association difference, and (B) signed semantic distance using direct associations as merit. Mark colors indicate endpoint colors. Error bars indicate standard error.

correlated with the proportion of times participants chose the darker side ($r(12) = .97, p < .001$). This correlation was stronger than the correlation for signed association difference reported above ($z = 1.96, p = .05$). The trail of points that plateaued in Fig. 8A now compress onto signed semantic distance values near -1 in Fig. 8B.

From this strong linear relation, one may suppose that only merit from direct associations is needed to simulate inferred mappings for colormap visualizations. But, Exp. 2 only included conditions in which direct associations and the dark-is-more bias *conflicted*, and Exp. 1 suggested that when they were *consistent*, the dark-is-more bias dominated regardless of association strength difference. To understand the relative contribution of these two potential sources of merit in assignment inference, it is necessary to model data sampled from multiple points along the full range of signed semantic distance (see Exp. 3).

Summary. As in Exp. 1, Exp. 2 showed that direct associations override the dark-is-more bias when the association difference between the light and dark colors was sufficiently large. The pattern of inferred mappings was strongly predicted by simulations of assignment inference using merit from direct associations (signed semantic distance).

6 Experiment 3

In Exp. 3, we developed and tested a new method to combine multiple (sometimes conflicting) sources of merit to simulate assignment inference (Fig. 1). The experimental task was the same as Exps. 1-2, but we tested three new domain concepts (ocean water, wild fire, and glacial ice), and sampled 21 points along the full range of direct association-based signed semantic distance for each concept. Our goals were: (1) determine the optimal weighting on direct associations (W_A) relative to the dark-is-more bias (W_D) when computing combined merit, and (2) test whether simulations of assignment inference using the optimal combined merit predicted people's inferred mappings better than simulations using each source of merit alone.

6.1 Methods

We first collected direct association ratings for the domain concepts and then used the mean ratings to generate colormaps to assess inferred mappings. We also collected additional data to quantify merit for direct associations and the dark-is-more bias.

6.1.1 Measuring direct color-concept associations

We collected direct association ratings for five domain concepts relevant to environmental data: wild fire, ocean water, glacial ice, ground soil, and tree foliage (inspired by [34]), using the same methods as in Exp. 1. The data are shown in Supplementary Material Fig. S.7.

The target sample was $n = 35$ to match Exp. 1, and we collected data in batches until reaching this target after excluding those with atypical 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.

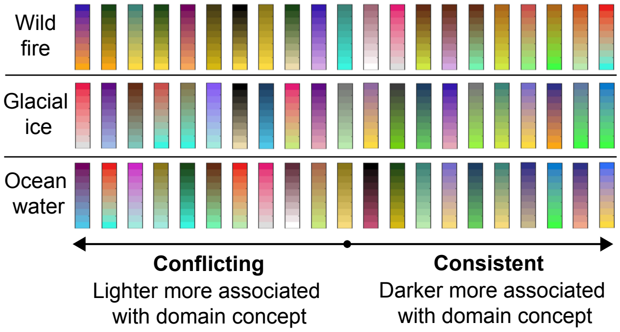
Wild fire
Glacial ice
Ocean water
Conflicting
Lighter more associated with domain concept
Consistent
Darker more associated with domain concept

Fig. 9: Color scales used to create colormaps in Exp. 3. Rightward of center, the domain concept increases in direct association strength with the *darker* endpoint (consistent direct and relational associations). Leftward of center, the domain concept becomes increasingly more associated with *lighter* (conflicting direct and relational associations).

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 × 21 color scales = 63 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_A) and the dark-is-more bias (W_D) 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 signed semantic distance and the mean probability that participants in the training set chose the darker side of the colormaps. When computing MSE, we scaled the proportion chosen data to range from -1 to 1, corresponding with the scale of signed semantic distance. Fig. 10A shows MSEs averaged over the 21 color scales for each domain concept, plotted as a function of weight pairs, along with the average over domain concepts. On average, the best performing weight pair yielding the lowest MSE had a weight of $W_A = .7$ on direct associations and $W_D = .3$ on the dark-is-more bias.

Using data from the held out testing set, we evaluated whether this optimal weight pair was better for predicting assignment inference than each source of merit alone. For each color scale for each domain concept, we computed MSE between mean responses (scaled to range from -1 to 1) and signed semantic distance with the optimal weighting identified from the training set (.7, .3), with all weight on direct associations (1,0), and with all weight on the dark-is-more bias (0,1) (Fig. 10B). To test effects of relative weighting, we used a linear mixed effects model predicting MSE for each color scale, with fixed effects for relative weighting, domain concept, and their interaction (using Helmert contrasts). The model also included a by-color scale random intercept and random slope for relative weighting. Here we focus on the main effect of relative weighting ($F(2,53.45) = 9.18, p < .001$), and we report on further details of this model in Supplementary Material Section S.7. Planned independent samples t-tests indicated that the optimal weight pair fit inferred mappings better than direct associations alone ($t(124) = -2.55, p = .01$) and dark-is-more alone ($t(124) = -3.53, p = .001$).

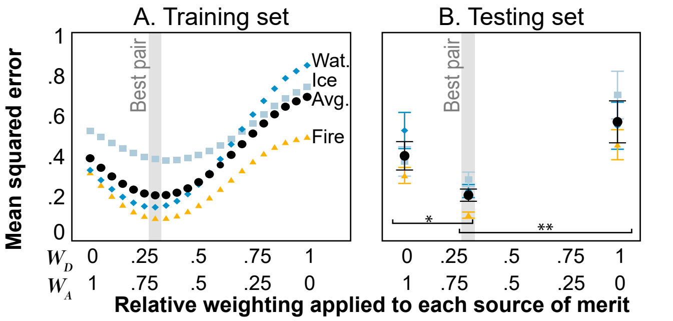

Fig. 10: Mean squared error (MSE) predicting inferred mappings from assignment simulations with varying weights on dark-is-more bias (W_D) and direct associations (W_A) for the (A) training set and (B) testing set. MSEs are shown separately for colormaps representing ocean water (blue diamonds), glacial ice (gray squares), and wildfire (yellow triangles), plus the average of all three concepts (black circles). Error bars represent standard error of the means. The gray bar indicates the best pair, determined from the training set.

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]

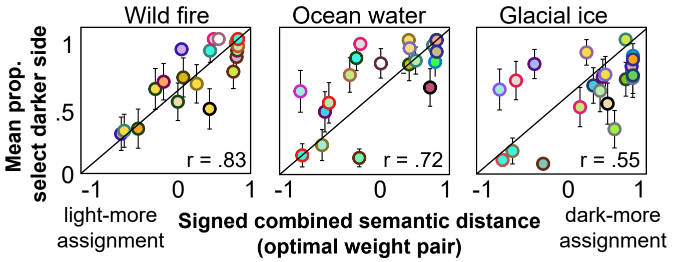

Fig. 11: Mean proportion of times the darker side was selected for the held-out testing set for each color scale as a function signed semantic distance using the optimal weight pair. Mark colors indicate color scale endpoint colors, and error bars indicate standard error.

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 plateau 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 endpoint 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 Zimnicki, 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.

REFERENCES

[1] F. M. Adams and C. E. Osgood. A cross-cultural study of the affective meanings of color. *Journal of Cross-Cultural Psychology*, 4(2):135–156, 1973.

[2] A. N. Bartel, K. J. Lande, J. Roos, and K. B. Schloss. A holey perspective on venn diagrams. *Cognitive Science*, 46(1):e13073, 2021.

[3] J. Blachowicz. Analog representation beyond mental imagery. *The Journal of Philosophy*, 94(2):55–84, 1997.

[4] 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.

[5] C. A. Brewer. Spectral schemes: Controversial color use on maps. *Cartography and Geographic Information Systems*, 24(4):203–220, 1997.

[6] R. Bujack, T. L. Turton, F. Samsel, C. Ware, D. H. Rogers, and J. Ahrens. The good, the bad, and the ugly: A theoretical framework for the assessment of continuous colormaps. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):923–933, 2018.

[7] M. Christen, D. A. Vitacco, L. Huber, J. Harboe, S. I. Fabrikant, and P. Brugger. Colorful brains: 14 years of display practice in functional neuroimaging. *NeuroImage*, 73:30–39, 2013.

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

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

[10] N. Elmqvist, P. Dragicevic, and J.-D. Fekete. Color lens: Adaptive color scale optimization for visual exploration. *IEEE Transactions on Visualization and Computer Graphics*, 17(6):795–807, 2010.

[11] G. P. Goodwin and P. Johnson-Laird. Reasoning about relations. *Psychological Review*, 112(2):468, 2005.

[12] 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.

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

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

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

[16] G. Kindlmann, E. Reinhard, and S. Creem. Face-based luminance matching for perceptual colormap generation. In *IEEE Visualization, 2002. VIS 2002.*, pages 299–306. IEEE, 2002.

[17] L. Kumle, M. L.-H. Võ, and D. Draschkow. Estimating power in (generalized) linear mixed models: An open introduction and tutorial in r. *Behavior Research Methods*, 53(6):2528–2543, 2021.

[18] 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.

[19] 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.

[20] C. J. Maley. Analog and digital, continuous and discrete. *Philosophical Studies*, 155(1):117–131, 2011.

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

[22] K. Mukherjee, B. Yin, B. E. Sherman, L. Lessard, and K. B. Schloss. Context matters: A theory of semantic discriminability for perceptual encoding systems. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):697–706, 2022.

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

[24] S. K. Murthy, T. L. Griffiths, and R. D. Hawkins. Shades of confusion: Lexical uncertainty modulates ad hoc coordination in an interactive communication task. *Cognition*, 225:105152, 2022.

[25] P. Nardini, M. Chen, F. Samsel, R. Bujack, M. Böttinger, and G. Scheuermann. The making of continuous colormaps. *IEEE Transactions on Visualization and Computer Graphics*, 27(6):3048–3063, 2021.

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

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

[28] L. Padilla, P. S. Quinan, M. Meyer, and S. H. Creem-Regehr. Evaluating the impact of binning 2d scalar fields. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):431–440, 2016.

[29] S. Palmer. *Fundamental aspects of cognitive representation. In Roach, E. and Lloyd, B. B., editors, Cognition and Categorization*. Lawrence Elbaum Associates, Hills- dale, NJ., 1978.

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

[31] 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.

[32] P. L. Rheingans. Task-based color scale design. In *28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making*, volume 3905, pages 35–44. International Society for Optics and Photonics, 2000.

[33] R. E. Roth, A. W. Woodruff, and Z. F. Johnson. Value-by-alpha maps: An alternative technique to the cartogram. *The Cartographic Journal*, 47(2):130–140, 2010.

[34] F. Samsel, T. L. Turton, P. Wolfram, and R. Bujack. Intuitive colormaps for environmental visualization. In *Proceedings of the Workshop on Visualisation in Environmental Sciences*, pages 55–59, 2017.

[35] 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.

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

[37] K. B. Schloss, L. Lessard, C. Racey, and A. C. Hurlbert. Modeling color preference using color space metrics. *Vision Research*, 151:99–116, 2018.

[38] 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.

[39] K. B. Schloss, C. Witzel, and L. Y. Lai. Blue hues don't bring the blues: questioning conventional notions of color–emotion associations. *Journal of the Optical Society of America A*, 37(5):813–824, 2020.

[40] G. D. Schott. Colored illustrations of the brain: some conceptual and contextual issues. *The Neuroscientist*, 16(5):508–518, 2010.

[41] 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.

[42] R. N. Shepard and S. Chipman. Second-order isomorphism of internal representations: Shapes of states. *Cognitive psychology*, 1(1):1–17, 1970.

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

[44] S. Silva, B. S. Santos, and J. Madeira. Using color in visualization: A survey. *Computers & Graphics*, 35(2):320–333, 2011.

[45] S. Smart, K. Wu, and D. A. Szafir. Color crafting: Automating the construction of designer quality color ramps. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1215–1225, 2019.

[46] C. Spence and G. Van Doorn. Visual communication via the design of food and beverage packaging. *Cognitive Research: Principles and Implications*, 7(1):1–23, 2022.

[47] 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.

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

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

[50] B. Tversky. Visualizing thought. *Topics in Cognitive Science*, 3:499–535, 2011.

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

[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.