Rainbows Revisited: Modeling Effective Colormap Design for Graphical Inference

Khairi Reda and Danielle Albers Szafir

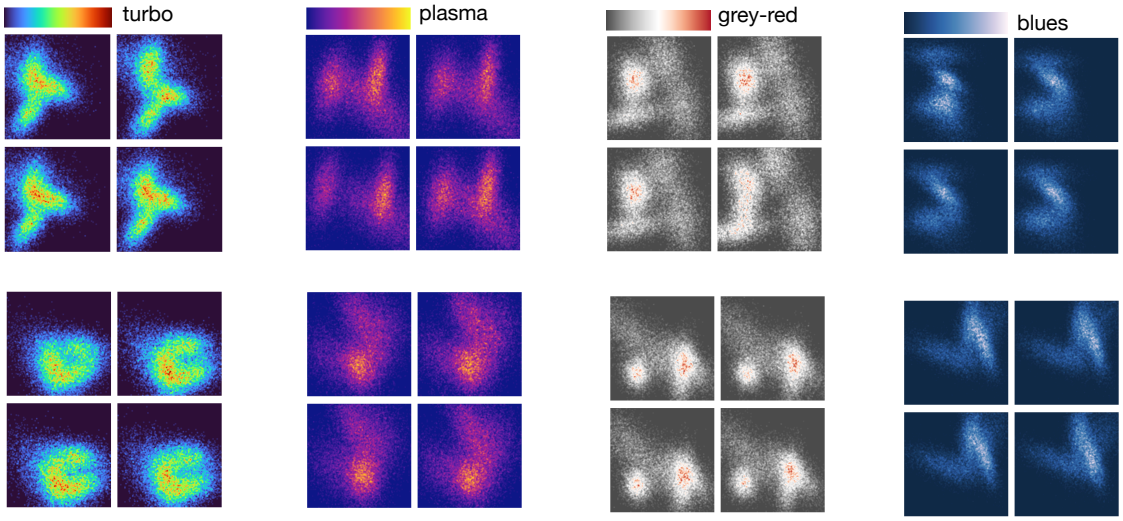

Fig. 1. Eight example stimuli from Experiment 1. A single stimulus consists of a lineup of four color-coded scalar fields shown in a 2×2 grid. For each lineup, which of the four plots stands out as different? The answers are in Section 10. This graphical inference test enables us to determine the discriminative power of competing colormap designs. Our results give rise a new model for predicting a colormap's usefulness, particularly for tasks involving model-based inference and judgement.

Abstract—Color mapping is a foundational technique for visualizing scalar data. Prior literature offers guidelines for effective colormap design, such as emphasizing luminance variation while limiting changes in hue. However, empirical studies of color are largely focused on perceptual tasks. This narrow focus inhibits our understanding of how generalizable these guidelines are, particularly to tasks like visual inference that require synthesis and judgement across multiple percepts. Furthermore, the emphasis on traditional ramp designs (e.g., sequential or diverging) may sideline other key metrics or design strategies. We study how a cognitive metric—*color name variation*—impacts people's ability to make model-based judgments. In two graphical inference experiments, participants saw a series of color-coded scalar fields sampled from different models and assessed the relationships between these models. Contrary to conventional guidelines, participants were more accurate when viewing colormaps that cross a variety of uniquely nameable colors. We modeled participants' performance using this metric and found that it provides a better fit to the experimental data than do existing design principles. Our findings indicate cognitive advantages for colorful maps like rainbow, which exhibit high color categorization, despite their traditionally undesirable perceptual properties. We also found no evidence that color categorization would lead observers to infer false data features. Our results provide empirically grounded metrics for predicting a colormap's performance and suggest alternative guidelines for designing new quantitative colormaps to support inference. The data and materials for this paper are available at: `https://osf.io/tck2r/`

Index Terms—Color, perception, graphical inference, scalar data

1 Introduction

Color is one of the most widely used channels in visualization. Designers use color to encode a variety of quantitative information, such as the wind-speed of a hurricane or the level of noise created by a jet engine simulation. In these and other similar visualizations, color not only communicates individual data values but also helps convey forms and patterns [59]. An observer can then study these patterns to infer something about the underlying physical process or model that generated the data.

Prior work offers a variety of guidelines for effective colormap design (see Bujack et al. [11] or Zhou & Hansen [68] for surveys). For example, experts advocate for color sequences that gradually increase in luminance for continuous variables [43, 51]. The idea behind this recommendation is that by carefully controlling luminance, we help establish ordinality. Additionally, luminance has higher capacity to convey subtle spatial details compared to hue or chroma [14, 61]. Experts have long discouraged the use of the *rainbow* and other *spectral* sequences, which tend to vary predominantly in hue [5, 35, 45]. Rainbow colormaps in particular have been singled out as an example of ineffective, or even deceptive, visualization design [5, 45].

The above recommendations have been absorbed into the canon of data visualization [36]. However, empirical studies of color driving

- *Khairi Reda is with Indiana University-Purdue University Indianapolis. E-mail: redak@iu.edu.*
- *Danielle Albers Szafir is with University of Colorado Boulder. E-mail: danielle.szafir@colorado.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

these guidelines have predominantly emphasized perceptual tasks, such as estimating values at specific map locations [41, 59] or comparing the distance between color patches [32]. Though informative, these studies do not always reflect the more complex and interpretive tasks people perform with visualizations, making the applicability of guidelines uncertain. Such tasks often require analysts to integrate multiple patterns and statistics to make inferences about visualized data. Consider a computational scientist who analyzes an ensemble of color-coded plots representing different model outputs. In this context, the scientist may be interested in understanding the models behind each visualization, comparing these models, and making a judgement about the relationships between models. The scientist may conclude, for instance, that the models are largely in agreement despite apparent variation in output or may draw on prior expertise to infer that a specific model shows a seemingly unusual outcome. In deciding what colormap to use for this kind of analysis, we could surmise that the component perceptual tasks, such as comparing key values, are best supported by a sequential or divergent ramp (e.g., *viridis* or *grey-red*). However, it is unclear if the more cognitive operations of model inference and judgement, such as distinguishing meaningful differences from noise, are similarly impacted. It is possible for cognitive determinants, such as the ability to distinguish categorically between colors or associate colors with distinct names and concepts [20, 31, 49], to play an important role in visual inference. A difference in colormap processing between the more cognitive versus perceptual tasks may necessitate different guidelines for effective visualization design, depending on the nature of the analysis.

Our goal in this work is two-fold. First, we aim to identify colormap design principles for improving graphical inference from quantitative visualizations. Second, we test whether certain colormaps can cause people to see false differences that are not present in the underlying data (i.e., false positives). To that end, we compare traditional colormap designs against an unconventional cognitive metric: ***color name variation***. We study the impact of this metric on people's ability to make graphical judgements about ensemble visualizations. Participants saw a lineup of scalar fields (Figure 1), and were prompted to identify an 'oddball' plot that belongs to an incompatible model. We found that colormap characteristics predictably affect performance at this task, with rainbow schemes affording the highest accuracy. Specifically, accuracy was positively correlated with a colormap's level of name variation, providing significant advantage to maps that traverse a variety of nameable colors. In a second experiment, we measured the impact of this metric on specificity to test the hypothesis that color categorization leads to false inferences. Results showed that, despite the increased sensitivity, rainbow maps did not necessarily lead to more false positives. Our findings collectively suggest benefits to maximizing categorical color and name variation in quantitative maps. We discuss the results, speculate on the underlying perceptual and cognitive mechanisms, and suggest alternative colormap design strategies.

2 Related Work

We review core color-encoding principles and discuss current approaches to colormap design. We then survey recent empirical studies of color in visualization, focusing on quantitative (as opposed to categorical) colormaps.

2.1 Color Mapping Guidelines

There is a rich body of guidelines for choosing color sequences for quantitative data [46, 51]. While the vocabulary varies from source to source [11], the guidelines generally agree on three principles [51]: 1) *Order*: a good colormap sequence should be naturally orderable (e.g., from a cool blue color to warm red); 2) *Continuity*: the colormap should only reflect actual differences in the data without creating artificial boundaries; and 3) *Perceptual Uniformity*: adjacent colors should reflect even perceptual distances throughout the sequence, such that a step in data magnitude is matched by an equivalent perceptual step in color.

Given these principles, most researchers advocate for ramps with monotonically increasing luminance [5] while discouraging the use of 'spectral' schemes (e.g., rainbows) [45]. Foundational textbooks also encourage the use of sequential and divergent maps for numeric data but recommend spectral schemes for categorical data [36]. The tendency for rainbows to create boundaries between hues (sometimes referred to as a 'hue banding' effect) is believed to mislead viewers [51]. However, the impact of such banding on data interpretation is still poorly understood [37, 39].

Despite consistent recommendations from researchers, rainbow colormaps continue to be popular among practitioners [35]. It is unclear why practitioners continue to adopt seemingly inferior designs. Researchers speculate that people find rainbow colormaps attractive [4], which in turn drives their color encoding choices. However, evidence also suggests that people prefer color combinations that are harmonious in hue [48], which most rainbows lack. Could practitioners' preferences reflect a utilitarian notion that is somehow missed by conventional guidelines? Our work considers this question in the context of a graphical inference task.

2.2 Approaches to Colormap Design

Tools for selecting or creating colormaps largely dictate the kinds of color encodings used in visualizations. Arguably, the most popular color selection tool is ColorBrewer [19], which provides a set of hand-crafted colormaps for quantitative or categorical variables. These colormaps are based on extensive empirical research [7, 8] and generally adhere to established guidelines. However, Brewer et al's color palettes are primarily aimed at thematic maps (i.e., choropleths) [6, 38]. It is unclear if these recommendations generalize to continuous spatial representations, such as scalar fields where data values typically create smooth color gradations.

Continuous maps appear frequently in domains such as computational science [66], medical imaging [4], astrophysics [29], and remote sensing for critical applications (e.g., visualizing hurricane data). The potentially large analytic and communicative impact for these visualizations highlights the need for validated color encoding principles. Although many practitioners continue to use rainbow colormaps, some visualization systems have adopted painstakingly crafted alternatives as defaults. For example, *cool-warm* in Paraview [34] and *viridis* in Matplotlib [57] aim to replace rainbow with more perceptually grounded alternatives (e.g., by ensuring perceptual uniformity). However, these systems offer a small library of fixed colormap options, making them difficult for designers to customize [31].

A few tools enable designers to construct their own colormaps guided by conventions. Designers can create ramps by specifying a handful of key points, which are then mapped to a geometric path (e.g., a line or simple curve) and sampled uniformly in a perceptual color space. For example, Wijffelaars et al. observed that ColorBrewer ramps tended to traverse cubic curves [65]. Their tool enables people to manipulate key control points on these curves, generating maps that mimic ColorBrewer. Other approaches rely on specific tasks to drive colormap selection [2, 56]. For example, PRAVDAColor [2] applies design conventions based on data characteristics and task types to generate tailored colormaps. However, more recent work suggests that assumptions around colormap design, such as the assertion that CIELAB matches perceived differences in data marks [52, 54], may not hold in practice. A recent analysis of practices also found that designer colormaps often do not obey conventional guidelines [53]. Empirical evaluations of colormap performance point to conflicting results that often cannot be reconciled with the guidelines [41, 42]. The limited empirical data, along with the near universal reliance on standard design tools, highlight the importance of validating color encoding guidelines in realistic tasks.

2.3 Empirical Studies of Quantitative Colormaps

Historically, colormap design guidelines have been largely grounded in intuition. Subsequent empirical studies sometimes confirm [32, 60] and sometimes challenge [41, 62] those guidelines. In an early study, Ware argued for colormaps that monotonically increase in luminance to aid form comprehension [59]. However, the same study also found non-monotonic variations (e.g., in rainbow) to reduce simultaneous contrast errors, effectively improving value estimation. Reda et al. confirmed

this hypothesis, but found the spatial frequency characteristics of the data largely dictate the optimal colormap design [41]. Ware et al. also evaluated how well viewers can find high spatial frequency features with different colormaps. They found that the detection threshold cannot be solely attributed to luminance, but is rather added by chroma and hue variation [62]. Some earlier work supports the effectiveness of rainbow schemes in interpreting thematic maps and continuous surfaces [7, 23, 30]. Recent studies, however, report findings that are in line with traditional preferences towards luminance-oriented colormaps. For example, in a study with doctors, Borkin et al. found rainbow colormaps to be substantially less accurate than divergent schemes for diagnosing heart disease [4]. The authors hypothesized that divergent ramps are better for finding low-value regions, which were key to completing that task accurately. Liu & Heer also found *jet* (a rainbow variant) to be generally inaccurate for judging perceptual distances between color patches [32]. Dasgupta et al. found that climate scientists' can more accurately estimate mean map values with a sequential scheme than with a rainbow [13]. However, they reported higher utility for rainbow maps in variance estimation. Using methods borrowed from psychophysics, Reda & Papka also found that people exhibit lower JND (i.e., higher sensitivity) when estimating spatial variance with rainbow [42].

In this work, we study a task that emphasizes graphical inference [10]. This kind of task allows us to test how well viewers can discriminate between models underlying a set of color-coded visualizations. We suspect such task to be important in many scientific contexts, where one may not only be interested in a single feature, but also in interpreting the relationships between multiple datasets. We also anticipate that such interpretive tasks require different colormap characteristics than those necessary for perceptual tasks, such as value estimation, based on evidence from prior studies. In this paper, we specifically compare perceptually based design recommendations against a cognitive metric of color categorization [20].

3 Methods

Visual analysis of 'ensembles' arises frequently in science. For example, climate scientists often analyze scalar field data to forecast global temperature rise under different climate models [66]. An aerospace engineer might be interested in comparing simulated noise levels generated by different jet engine prototypes [27]. Beyond simulated data, infectious disease researchers often study color-coded 2D histograms to understand the receptor characteristics of immune cell populations subjected to flow cytometry techniques [50]. These tasks are not limited to decoding raw values from color. Rather, analysts often need to infer properties of the model behind the visualizations and potentially assess the relationships between multiple alternative models. Inference is further complicated by the fact that a visualization typically represents a single sample from what could be a large (or infinite) number of potential samples of either a model or the real world. One may obtain a visually distinct visualization of the same phenomenon by simply re-sampling the model or by taking a new set of observations. An analyst therefore has to think about which image features are due to real model properties and which are due to random variations arising from the sampling process. The complexity involved in this type of task provides an opportunity to probe our assumptions about how color works in visualization, while also examining an important class of analyses relevant to many practitioners (e.g., professional scientists).

3.1 Model Task

To simulate the above analyses experimentally, we devised a task based on Buja et al's concept of graphical inference [10, 63]. The task evaluates how well a viewer can visually discriminate between two distributions. The idea is to conceal a plot of the 'real' data, which typically contains a pattern of interest, in a lineup that also contains several 'decoy' visualizations. The decoy plots represent datasets that could have arisen by chance, typically sampled from a null distribution. The viewer's task is to identify the real data from the null by inferring true differences between the two distributions.

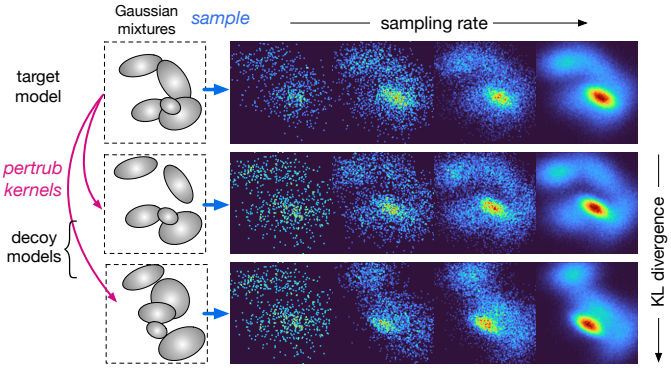

Fig. 2. Stimulus scalar fields are synthesized by iteratively sampling from a Gaussian mixture model. The first row corresponds to a target model. The second and third rows show two example decoy models obtained by perturbing the target. The third row corresponds to a decoy with larger perturbation (higher KL divergence) for an easier trial. The columns represent increasing sampling rates, with the rightmost column almost perfectly recreating the actual probability density. In our experiments, sampling is stopped near Column 3 to provide participants with an imperfect representation of the model.

Buja et al's protocol was conceived as a method for guarding against false discovery [10]. However, the test is also useful for evaluating the power of competing visualization techniques [12, 22, 64]: if a particular visualization method makes it easier to distinguish smaller model differences, then that visualization can be said to have higher statistical power. For example, Hofmann et al. employed graphical inference to compare two different visualization designs for timeseries [22]. We use a similar approach to measure the discriminative power of different colormap designs for inference. We focus on continuous colormaps for scalar data and 2D histograms as those visualizations arise frequently in scientific applications, where we often see a need for inference. In our experiments, the viewer sees a lineup of four color-coded scalar fields (see Figure 1 for examples). The fields are sampled from one of two models: a *target* and a *decoy*. Specifically, one of the four visualizations is sampled from the target model, while the other three are sampled from the decoy model. Per the original method [63], the viewer must identify the plot that "doesn't belong:" in this case, the visualization that came from the target model.

Graphical inference can be seen as analogous to a statistical hypothesis test in that both are intended to discriminate true differences [10, 63]. However, it is important to recognize that visualizations support a variety of inferential styles, and the task we adopt here represents one instance from a large class of inferential tasks. In this paper, we use the term 'inference' as a shorthand for graphical inference.

3.2 Stimuli

We employ a synthetic data generation procedure to generate 2D fields with known target-decoy model pairs. The procedure creates a target model and subsequently perturbs it to produce decoys (see Figure 2 for an illustration of this process). The target is first synthesized from a combination of 2D Gaussian kernels. The kernels are centered randomly within the spatial domain and their parameters (expected value, standard deviation, and covariance) are varied to generate pseudorandom configurations. The kernels are then integrated to produce a joint 2D probability distribution. To generate a decoy model, individual kernel parameters from the target model are randomly perturbed within a fixed range (determined through piloting). Larger perturbation results in more obvious target-decoy divergence and, thus, easier judgement. This entire process is repeated separately for every trial.

The synthetic procedure above produces models that exhibit a variety of spatial arrangements and patterns, as shown in Figures 1 and 2. Qualitatively, our implementation generates both models dominated by large densities as well as those that are composed of smaller features.

The emerging target-decoy pairs also differ in a variety of ways across stimuli; we observe differences characterized by shifting of the densities or "empty spaces," addition or removal of "hotspots," and overall changes in the 2D pattern. This broad diversity helps ensure that our experimental results are not biased by specific shapes or patterns.

We generate a lineup stimulus (Figure 1) by sampling the target and decoy models to produce four fields (one from the target and three from the decoy). Because fields represent random draws from either distribution, a lineup will always exhibit variations between fields even for fields drawn from the same distribution. Visualizing samples, as opposed to the actual probability densities, provides participants with imperfect model representations, the quality of which is afflicted by the sampling process. This uncertainty is key to operationalizing graphical inference. First, the randomness introduced by sampling prevents participants from basing their judgements on small variations in one part of the image. Instead, they will need to compare the four visualizations holistically and integrate multiple percepts [40] to make a summary determination. Second, participants will need to distinguish visual features that reflect systematic model differences from fluctuations due to random sampling. The combination of these two factors give rise to a more interpretive task than in earlier experiments, which emphasize statistical or localized percepts [32,41,59,62]. This task in turn enables us to shift the focus from issues of color perception to questions of how color encodes affect inference and model-based judgement.

4 Hypotheses & Metrics

Color mapping guidelines, such as ensuring equal perceptual distances and minimizing color discontinuities, are inspired by traditional color spaces (e.g., CIELAB or LUV). These color models approximate the relative appearance of a small number of isolated color patches. By contrast, visualizations often comprise a significant number of marks or color gradations (as in scalar fields or heatmaps). The layouts and viewing conditions in these data displays are far more complex than those assumed in conventional color models, making these models (and the guidelines on which they are based) less dependable in practice [54]. Furthermore, while metrics of perceptual color distance may be important in perceptual tasks (e.g., estimating a color value), those same metrics may arguably be less relevant for graphical inference, where precise quantitative differences may not matter as much as the overall quality of those differences.

We hypothesize that accessible cognitive characteristics of color [15], such as the ability to readily recognize colors by name and distinguish different hues categorically, play a larger role in graphical inference. In particular, we expect that color nameability aids people in reasoning over coarse-grained differences despite a potential loss of precision in comparing smaller perceptual variations. Recent empirical work provides evidence to substantiate this conjecture. For example, Reda & Papka found that observers efficiently estimated structural properties (e.g., gradients) in scalar fields using rainbows [42]. Participants appeared to take advantage of the emerging discrete color bands (e.g., blue, green, and red patches in rainbow). By visually judging the apparent size, numeracy, and clusteringness of these features, one can obtain fairly reliable estimates about various statistical properties. Color categorization may seem like a bad idea for quantitative datasets [5], but this discretization may help in practice by providing a "featurized" representation of the data. Segmented features, in turn, can make it easier to extract certain ensemble properties at a glance [16].

Dogmatic adherence to "bad" colormaps by visualization practitioners may also suggest some unobserved utility. Many professional scientists continue to shun expert guidelines in favor of seemingly inferior rainbow maps [35,45]. Common justifications for this preference include that practitioners are hesitant to change their conventions [34] or find colorful visualizations to be aesthetically appealing [4]. However, it is also possible that practitioners find rainbows useful for some tasks, likely due to cognitive benefits that cannot be discerned from traditional color appearance models—a hypothesis that we test. We also consider a counterargument that more colorful maps can cause people to "hallucinate" features that are not in the actual data [28]. The latter could manifest as a reduction in inference specificity.

We posit two hypotheses based on the above observations:

H1—We argue that perceptually grounded colormap metrics, such as uniformity, order, and smoothness [11], are not well-suited for inferential tasks. Instead, we hypothesize that performance will be tied to people's ability to reason categorically about color in a visualization. The latter may be facilitated by maps that blend a broader variety of nameable colors. Because there is no established way to measure the degree of categorical separation in continuous colormaps, we propose a new metric, *color name variation* (see §4.1) based on a popular name-distance model [20].

H2—We expect colormaps with high name variation to cause people to detect false differences between visualizations of the same model.

Color categorization is thought to be misleading for data types that vary continuously [5]. For example, the transition from green to yellow in a rainbow causes people to perceive a sharp boundary, which could be mistaken for a data feature. Therefore, while we anticipate people to be more sensitive with high name-varying colormaps, we suspect these ramps will also increase the rate of false positive judgments. The latter manifests as greater likelihood of reporting differences between visualizations when the underlying models are in fact identical.

4.1 Color Name Variation

Color names refer to the basic linguistic associations we make with color (e.g., "red," "green," and "blue,") [3,9,67]. Color naming can also be a cognitive tool, providing people with a way to categorize, discuss, and reason over color. In visualizations, color categorization may help people think about complex patterns by enabling selective attention to specific data features, which may otherwise be too fuzzy. Categorization can also help reduce a large number of data points to a smaller set of "bins" that can be reasoned with more easily. Therefore, the extent to which a colormap facilitates categorization may correlate with performance on graphical inference tasks (H1).

To approximate categorization, we use Heer & Stone's name distance model [20]: given two CIELAB colors, the model outputs the probability the pair have different names. We extend this metric to a continuous colormap by measuring the probability that an adjacent pair of colors are associated with distinct names. Summing these probabilities gives us an approximation of the number of distinct colors contained within the colormap. We refer to this measure as *color name variation* (CNV). Formally:

$$CNV(C) = \sum_{i=1}^{n} \Delta(C_{(i-1)/n}, C_{i/n}) \tag{1}$$

where $C_{i/n}$ is a color sampled from a continuous ramp C at ith position, n is the number samples to be taken (we use $n = 8$), and Δ is the cosine name-distance from Heer & Stone [20], computed as:

$$\Delta(C_{(i-1)/n}, C_{i/n}) = 1 - cos(T_{C_{(i-1)/n}}, T_{C_{(i)/n}}) \tag{2}$$

where T_C is the color-term count matrix derived from five million samples through an online survey.[1]

Higher name variation indicates a ramp that combines a variety of distinctly named colors, which in turn presents as an increased likelihood of color categorization. For example, rainbow colormaps which sweep through a range of saturated hues score higher on this metric, whereas single-hue ramps generally exhibit much lower name variation. Divergent and multi-hue ramps tend to overlap and lie somewhere between rainbows and single-hue ramps (see Figure 3). Although we intuitively associate color names with different hues, name variation is not uniquely bound to any single perceptual factor. Across a corpus of 235 colormaps (see §4.2), the CIE L^*, C^*, and h^* components exhibited comparable correlation with name variation (Pearson's $r = 0.754, 0.727$, and 0.718, respectively). Name variation thus appears to be almost equally driven by changes in lightness, chroma, and hue.

[1] http://blog.xkcd.com/2010/05/03/color-survey-results/

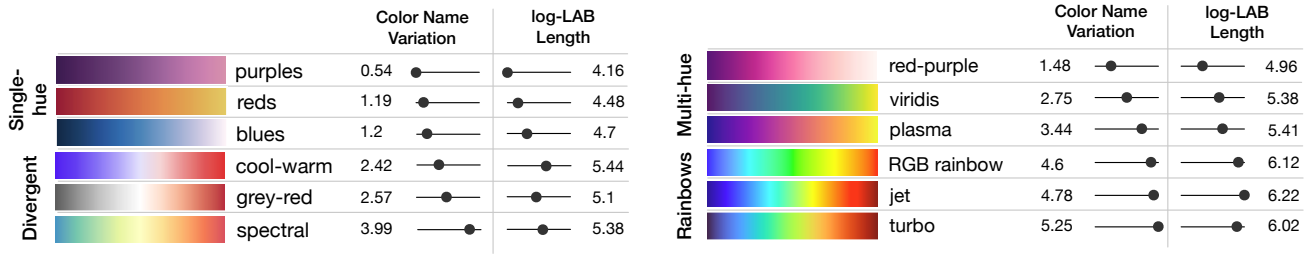

Fig. 3. We selected a set of 12 color ramps representing four conventional design families: single-hue, multi-hue, divergent, and rainbow ramps. These ramps exhibit a range of *color name variation* scores, the primary metric we use to model subject performance. We also consider a log-transformed CIELAB metric as an approximation of name variation.

4.2 Ramp Selection

To understand what colormap properties facilitate graphical inference, we sampled a set of 12 colormaps from a corpus of 235 unique designs. The original corpus included a variety of sequences collected from ColorBrewer [19], ColourLovers,[2] MATLAB, Matplotlib [57], Google [33], and Tableau. The resulting sample of 12 ramps represents a continuum of name variation levels across four different design families (single-hue sequential, multi-hue sequential, divergent, and rainbows). Figure 3 illustrates the selected ramps.

In addition to broadly representing commonly used colormaps, the selected ramps afford interesting comparisons. For example, ColorBrewer's *red-purple* has higher name variation than ColourLover's *purples* but lower than *plasma*, yet the three ramps share a similar gamut of purple hues. We similarly included three rainbow variants: *turbo*, *jet*, and a standard *RGB rainbow*. This selection represents relatively high color name variation but also notable differences in perceptual characteristics. For example, Google's *turbo* is purported as a perceptually improved alternative to *jet* due to its almost-uniform luminance profile.

5 Experiment 1

Experiment 1 tests the hypothesis that a colormap's usefulness for graphical inference is correlated with its name variation (H1). We expect participants to be more accurate with ramps that combine a larger variety of nameable colors. To test this hypothesis, we conducted a crowdsourced study with colormap as a between-subjects factor.

We analyze the results using three models, each providing a competing explanation of the results. The first model predicts that colormaps from the same **design family** will show comparable accuracy but that the different families (single-hue, multi-hue, divergent, and rainbow) will exhibit varying performance. By contrast, the second model is more parsimonious and predicts performance using a single quantitative metric: the colormap's **name variation**. The model is able to distinguish more granularly between colormaps, even for those within the same design family (e.g., *viridis* and *plasma*). The third model uses a colormap's log-transformed **length in CIELAB** space as an approximation for name variation.

By comparing these alternative models, we assess the suitability of our two metrics (*color name variation* and *log-LAB length*) for predicting effective colormap composition. Additionally, by contrasting with the *design family* model, we test if these metrics provide a more useful indicator than conventional guidelines (e.g., perceptual uniformity and luminance monotonicity). We use the Bayesian Information Criterion (BIC) to compare model fit. The choice of BIC reflects our goal in uncovering general colormap design principles as opposed to accurately predicting performance on a per-subject basis. BIC favors simpler models by penalizing the number of parameters; however, more complex models are still favored if they improve the fit to the data.

[2]https://www.colourlovers.com

5.1 Task

We employ the graphical inference task described in §3.1. Each stimulus comprises a lineup of four scalar fields (see Figure 1) that are sampled from one of two models: a target and a decoy model. All generated fields measured 200×200 pixels, subtending approximately 4°of visual angle when viewed from 30 inches away at 96 DPI (the standard pixel density for web browsers [58]). A color scale was displayed to the right of the lineup for reference. Participants were instructed to identify the visualization that "does not belong." They indicated their choice by clicking on one of the four images and confirmed the selection by pressing Enter to move to the next trial.

We limited lineups to four visualizations, as opposed to the more typical ten plots [63], to reduce the per-trial response time and allow for a larger number of stimuli per subject. While this setup increases the alpha-level for an individual lineup test from 0.1 to 0.25, our experimental design does not depend on reaching statistical significance in every trial. Instead, to accommodate the reduced statistical power, we increase our sampling rate and collect 88 trials per participant. We also employ a diverse set of distributions to ensure the results are not biased by specific model features (e.g., clusters or hotspots).

5.2 Experimental Design

We evaluated 12 colormaps (Figure 3, §4.2) in a between-subjects design. Every participant saw one colormap from each of the four design families for a total of four colormaps per participant.

The experiment was blocked by colormap with block order randomized. Participants completed 22 trials with each colormap, resulting in 88 trials per participant. Each trial consisted of a freshly generated lineup (as described in §3.2). Difficulty was controlled by systematically varying the divergence between the target and decoy models. We measured divergence using Kullback-Leibler's (KL), a popular information-theoretic metric for quantifying the distance between two distributions [21]. We uniformly sampled stimuli with KL divergence in a range of 10–35%. Lower divergence corresponds to smaller model differences (i.e., more difficult judgements). This range was determined through piloting to reflect expected success probabilities that are slightly greater than chance ($P = 0.25$) to near perfect. For a visual reference, the top row in Figure 1 represents stimuli with 27% divergence (somewhat easy) whereas the bottom row is at 15% (less obvious differences).

Individual trials within each block (corresponding to different levels of difficulty) were displayed in random order. In addition to the actual trials, we randomly inserted 2 engagement checks per block for a total of 8 checks throughout the experiment. These checks consisted of very easy stimuli (40% divergence, or four times easier than the starting difficulty).

5.3 Procedure

Participants were first screened for color-vision deficiency using 14 Ishihara panels. They then saw a tutorial and completed 24 practice trials. These practice trials included feedback informing participants of whether they had guessed correctly or indicating the correct answer otherwise. During training, each participant saw their four assigned

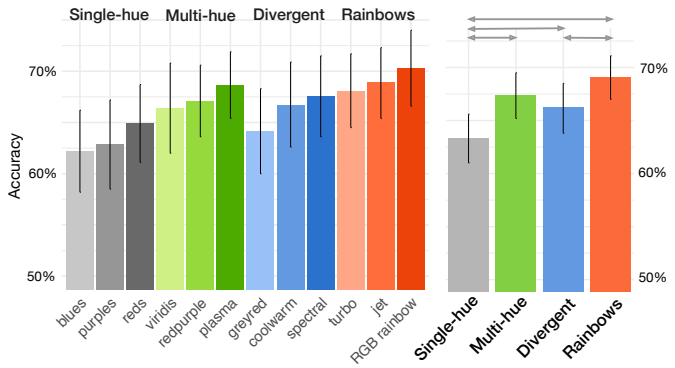

Fig. 4. Observed mean accuracy for the 12 colormaps (left). The same data is also grouped by design family (right). Errors bar are 95% confidence intervals. Arrows indicate significant differences in accuracy between the design families ($p < 0.05$ with Tukey's adjustment).

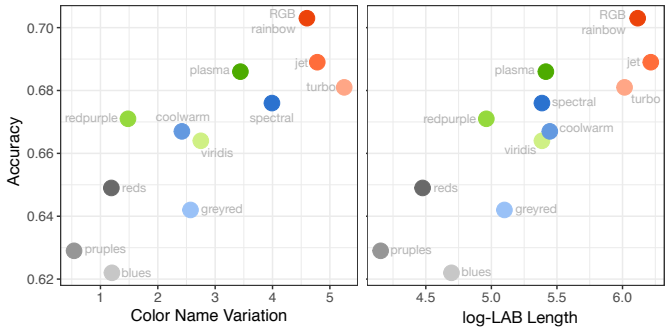

Fig. 5. Mean colormap accuracy as a function of *color name variation* (left) and *log-LAB length*. Both metrics show significant correlation with the observed inference performance in Experiment 1.

colormaps in random order. After practice, participants completed the actual trials, in which no feedback was given. At the end of the experiment, we asked participants to provide a brief qualitative description of the strategy they followed in the task including any "visual features or characteristics" they based their judgements on. Participants concluded by completing a brief demographic survey.

5.4 Participants

We recruited 180 participants (115 males, 64 females, and 2 others) from Amazon Mechanical Turk, compensating them with $3 each. Participants had a mean self-reported age of 37.4 years ($STD = 11.7$). We excluded from the analysis any participant who failed the color-vision test or misjudged the majority of the engagement checks (mean performance on the engagement checks was 95.6%). We then recruited new participants to replace those excluded until we reached the sample size above. To mitigate effects of interparticipant variation, every colormap was seen by exactly 60 different individuals, and each colormap from a particular design family was tested at least twice with every other colormap from a different family.

5.5 Results

Participants completed the experiment in 20.4 minutes on average ($STD = 8.2$). In total, we obtained 15,840 binary judgements (i.e., whether the participant had correctly identified the target distribution). Mean accuracy at the task was 66.5% (chance performance is 25%). Figure 4 plots accuracy as a function of colormap and design family.

We analyze the results using three logistic regression models: 1) a model predicting the probability of correct inference separately for each **design family**; 2) a model that predicts performance based on a colormap's **name variation**; and 3) a model representing *post-hoc* refinement of the second model with **log-LAB length** serving as an approximation to name variation (raw LAB distance resulted in poor fit that we were unable to obtain a well-formed model). All three models included one additional fixed-effect parameter corresponding to the target-decoy divergence (i.e., trial difficulty) and one random intercept to account for individual variation among subjects. We first analyze each model separately and then compare their relative fit. Figure 6 plots model responses against the observed data. Table 2 gives the fitted parameters for each model.

5.5.1 Design Family

The design family model contains four discrete variables, one for each of the four design groups (single-hue sequential, multi-hue sequential, divergent, and rainbows). On average, participants were most accurate when viewing rainbow colormaps (Mean accuracy: 69.1%, 95% CI: 67.0–71.1). Multi-hues were the next most effective (mean: 67.4%, CI: 65.2–69.5) followed by divergent (mean: 66.2%, CI: 63.8–68.5) and single-hue ramps (mean: 63.3%, CI: 61.0–65.6).

We draw pairwise comparisons between the four design families using Wald's Z and employ Tukey's adjustment for multiple comparisons (see Figure 4-right for a summary of the significant differences). Rainbow colormaps were significantly more accurate than single-hue ($Z = 6.020, p < .001$) or divergent ramps ($Z = 3.107, p < 0.05$). Multi-hue and divergent colormaps were both more accurate than single-hues ($Z_{mh} = 4.153, p_{mh} < 0.001$; $Z_d = 2.922, p_d < 0.05$).

5.5.2 Color Name Variation

Our second model captures performance solely as a function of color name variation, making no distinction between design families. A Wald's test indicates name variation to be a significant predictor of performance ($Z = 5.526, p < 0.001$). A step increase in this parameter improves the odds of correct inference by a factor of 1.07. In other words, color name variation is positively correlated with accuracy. This translates to sizable advantage for ramps blending a variety of distinctly nameable colors. As an example, the difference in name variation between *viridis* and *jet* stands at 2.03 in favor of the latter. Accordingly, the odd of inferring the correct model is 1.15 times higher with *jet* than with *viridis*, all other factors being equal.

5.5.3 log-LAB Length (*post-hoc*)

Color name variation can be cumbersome to compute: it is not supported by most design tools and requires access to an empirical name model. We therefore looked for a closely related metric that is grounded in more standard color spaces. We found that by taking the full length of a colormap's curve in the LAB space and log-transforming that measurement, we obtain a close approximation to name variation. For the colormaps included in this study, the two measures exhibit high correlation (Pearson's $r = 0.946$). Table 2-C lists coefficients for a model trained with this metric. The model indicates log-LAB length as a significant predictor of accuracy ($Z = 5.99, p < 0.001$): a step increase in this parameter improves the odds of correct inference by a factor of 1.2. For example, comparing *jet* and *viridis*, the model predicts inference odds that are 1.17 times higher with *jet*.

5.5.4 Model Comparison & Discussion

We use BIC scores to determine which of the models is a better fit to the data (see Table 1). Lower BIC indicates a more desirable model based on a trade-off between fit and parsimony. A difference greater than 10 provides strong evidence in favor of the lower scoring model [25]. The

Model	Parameters	AIC	BIC	logLik	Deviance
Design Family	5	17209.39	**17255.4**	-8598.7	17197.4
Color Name	3	17213.24	**17243.9**	-8602.6	17205.2
log-LAB Length	3	17207.92	**17238.6**	-8600.0	17199.9

Table 1. Comparison of the goodness-of-fit for the three models. To select among the models, we use the BIC criterion (bold, lower is better).

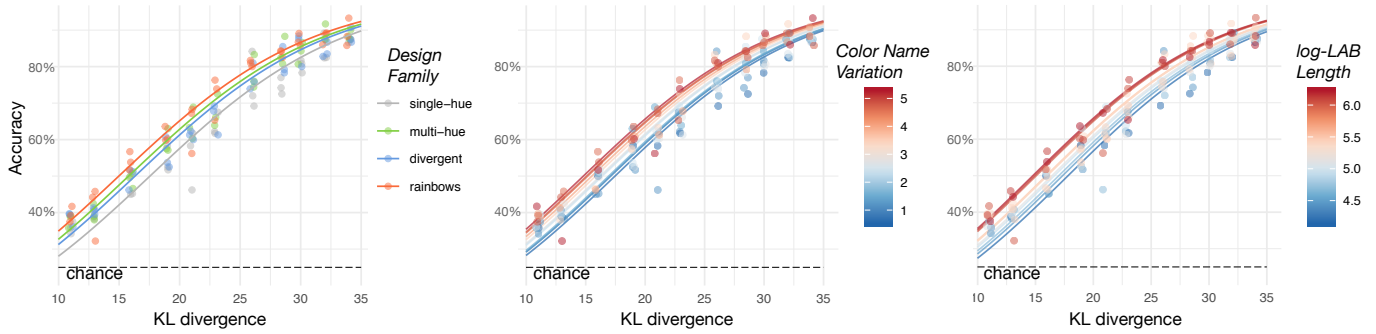

Fig. 6. The three models evaluated in Experiment 1 and their fit. Curves show model predictions. Colored dots depict mean subject accuracy for a particular divergence (i.e., difficulty) level.

Parameter	Estimate (95% CIs)	Z value	p
(Intercept)	0.39 (0.34−0.44)	14.18	***
divergence	1.13 (1.13−1.14)	45.33	***
Multi-hue	1.25 (1.12−1.38)	4.15	***
Divergent	1.17 (1.05−1.29)	2.92	**
Rainbows	1.38 (1.24−1.53)	6.02	***

(A) Design Family model

Parameter	Estimate (95% CIs)	Z value	p
(Intercept)	0.39 (0.35−0.45)	14.31	***
divergence	1.13 (1.13−1.14)	45.33	***
Color Name Variation	1.07 (1.05−1.1)	5.53	***

(B) Color Name model

Parameter	Estimate (95% CIs)	Z value	p
(Intercept)	0.38 (0.33−0.43)	14.42	***
divergence	1.13 (1.13−1.14)	45.33	***
log-LAB Length	1.20 (1.13−1.28)	5.99	***

(C) log-LAB Length model

Table 2. Parameters for the three models. The estimates shown correspond to exponentiated model coefficients so as to reflect odd ratios (±95% confidence intervals). Asterisks denote p-values (***=$p < 0.001$, **=$p < 0.01$)

BIC score for the *color name* model is lower than the *design family* model (ΔBIC=11.5). This difference corresponds to strong evidence that name variation is a better explanation of the empirical results. The *log-LAB length* model has a slightly lower BIC score than the original *color name* model it approximates, suggesting a slightly better fit; however, the difference is small (ΔBIC=5.3). The log-length of a colormap is therefore a good approximation for its name variation, and by extension, inference performance.

Our results show *name variation* and *log-LAB length* to be better predictors of colormap usefulness for graphical inference. Either metric alone appears to closely predict participants' expected performance (see Figure 5). Furthermore, the three models are in agreement and show that accuracy is bolstered by incorporating distinctly nameable colors. The correlation between name variation and accuracy coincides with a preference for more colorful ramps (e.g., *RGB rainbow* and *jet*), which afforded greater accuracy. These results run counter to conventional design wisdom, which stipulates that rainbows are ill-suited for quantitative data [5] and should instead be reserved for categorical variables [36]. Rather, we find that rainbow colormaps significantly outperform the more perceptually uniform alternatives, such single-hue (e.g., *blues*) and divergent ramps (e.g., *cool-warm* and *grey-red*). These counterintuitive results are explainable by considering differences in name variation.

6 Experiment 2

Results from Experiment 1 indicate that name variation improves graphical inference. Participants appear to benefit from colormaps that blend a variety of color names. However, the more colorful ramps (e.g., *jet* and *RGB rainbow*) may also be deceptive [43,44] as they implicitly discretize visualizations even when the underlying data is continuous [39]. The resulting visual artifacts can be misconstrued as data features [5], leading to false inferences. Experiment 2 tests the hypothesis that rainbow colormaps increase the rate of false discovery (H2). Specifically, we study how color name variation impacts the two kinds of error people make in inference: Type I (false positives) and Type II errors (false negatives).

6.1 Task

We modify the original graphical inference task (see §3.1) to model a situation where an analyst can either declare a positive (i.e., there is a detectable difference between a set of visualizations) or a negative result (there is *not* a difference, with the visualizations representing the same phenomenon). We adapted the Experiment 1 task by converting the response format from multiple choice (i.e., choose one of four plots) to a binary response (i.e., do these plots represent the same model). Participants indicated which of four visualizations in a lineup appears to come from a different model (a positive inference) or, alternatively, declared all four visualizations belong to the same model (a negative). This design is analogous to a dichotomous choice between either rejecting the null hypothesis by asserting that one of the visualizations in the lineup is special or accepting the null hypothesis and declaring that the visualizations correspond to the same distribution despite some noise.

We modify the sampling procedure (§3.2) such that, in half of the trials, the lineup plots are sampled from the same target model (i.e., no decoys). The other half consists of trials that are identical to the original task, with three of four visualizations sampled from the decoy model and one from the target. To discourage random guessing, we maintain a similar response format and prompt participants to click on the image that "does not belong". However, we add a fifth choice labeled "no discernible difference between the images." Our analysis treats the response as binary: positive if the participant selects one of the four images or negative for declaring a 'no difference.' False positives occur when participants falsely report a difference, such as those caused by hallucinated features from rainbow banding, whereas false negatives occur when participants fail to report differences, such as when model-differentiating features are undetected.

6.2 Experimental Design & Procedures

We selected a subset of four colormaps for this experiment: *blues*, *viridis*, *cool-warm*, and *RGB rainbow*. These ramps correspond to different levels of name variation (*blues*–low, *viridis* & *cool-warm*–mid, *RGB rainbow*–high) while also representing the four design families. In Experiment 1, these ramps exhibited near-linear performance differences as a function of name variation (Pearson's $r = .986$), providing a good source of error variation. We employ a within-subjects design, with all participants experiencing the four colormaps.

The procedures were similar to Experiment 1 (see §5.3 for details). Participants were first screened for color-vision deficiency. They then completed a practice session with feedback followed by four blocks of analyzed trials. The experiment was blocked by colormap with 24

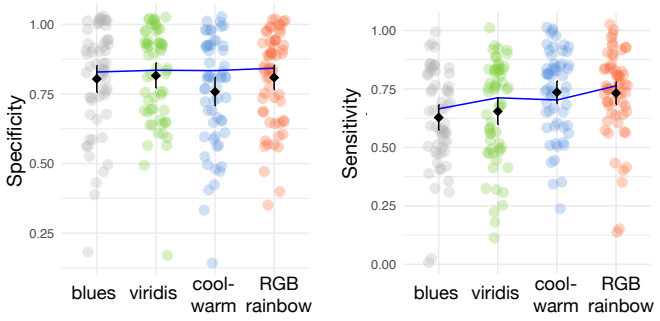

Fig. 7. Specificity and sensitivity by colormap. Colored dots depict rates for individual participants. Diamonds depict group means (± 95% CIs). Blue lines represent model predictions. A flat line (left) indicates that color name variation does not predict specificity; however, it is correlated with sensitivity.

trials per block (half true positives and half true negatives), for a total of 96 trials plus 8 engagement checks. Trial and block order were both randomized. To help clarify task instructions, we explicitly told participants that half of the stimuli will show no discernible difference between the images. This information was emphasized to reduce potential response bias due incorrect priors (e.g., participants wrongly assuming negative stimuli are more likely).

6.3 Results

We recruited 60 participants (mean age: 37.4 years, *STD*: 10) from Amazon Mechanical Turk, compensating them with $3 each. We used the same exclusion criteria as in Experiment 1 (see §5.4). Participants completed the experiment in 22.4 minutes on average (*STD*: 10.7). Mean accuracy in the task was 74.2%, with 68.8% sensitivity (true positive rate) and 79.7% specificity (true negative rate). Figure 7 plots the results by colormap. We employ logistic regression to separately analyze the specificity and sensitivity results. As in Experiment 1, we fit two models: a *color name* model and a *design family* model. We omit the *log-LAB length* model for space constraints and given its similarity to the name model.

6.3.1 Specificity

Color name variation does not predict specificity (exponentiated estimate: 1.03, CI: 0.95–1.12, Wald's $Z = 0.732, p = 0.464$). The estimate, centered approximately around one, suggests that colorful maps do not necessarily increase the odds of making a false positive error. However, the design family model shows some performance differences: the divergent *cool-warm* ramp exhibited significantly lower specificity (*ee:* 0.74, CI: 0.57–0.97, $Z = 2.212, p < 0.05$). Pairwise comparisons with Tukey's adjustment show *cool-warm* to be worse than *viridis* ($Z = 2.857, p < 0.05$). All other comparisons were not significant.

6.3.2 Sensitivity

Color name variation significantly predicts sensitivity (*ee*: 1.15, CI: 1.07–1.24, $Z = 3.951, p < 0.001$): a step increase in name distance increases the odds of resolving true positives by 1.15 times. Similarly, the design family model indicates a preference for rainbow and divergent designs: both *RGB rainbow* and *cool-warm* led to higher sensitivity than *viridis* ($Z_r = 3.451, p_r < 0.01$; $Z_{cw} = 3.579, p_{cw} < 0.01$) and *blues* ($Z_r = 4.428, p_r < 0.001$; $Z_{cw} = 4.554, p_{cw} < 0.001$).

In summary, the results provide no evidence that the more colorful maps (e.g., *RGB rainbow*) increase the false positive rate. We instead find evidence confirming that name variation supports inference by increasing the true positive rate. The higher sensitivity for *RGB rainbow*, combined with its baseline specificity, may explain why this colormap performed highly in Experiment 1. People appear better at discriminating between models when viewing a rainbow map without necessarily being misled into seeing false, model-extrinsic differences.

7 Discussion

Color is one of the most commonly utilized visual properties for communicating quantitative data. However, we are still developing an empirical understanding of how color works in visualization. We sought to understand how colormap characteristics affect graphical inference. We hypothesized that cognitive determinants, such as the ability to name and think categorically about colors, affect people's ability to draw inferences about data. We further hypothesized that inferential accuracy can be improved by using colormaps that cross a variety of color names. To quantify the latter, we proposed a new metric, *color name variation*, and studied its impact in two crowdsourced experiments.

7.1 Inference by Color

We first sought to model participants' performance by using a colormap's name variation as an explanatory factor. Results show name variation to be a good predictor of colormap utility for graphical inference. Incorporating a larger span of nameable colors significantly increases accuracy. This positive correlation can be seen in Figure 5-left. These results support **H1**.

Our results are at odds with conventional guidelines, which suggest that designers should limit colors to a judicious selection of hues [32] while emphasizing luminance variation and perceptual uniformity [26, 43, 62]. We instead found that the highest performing colormaps were rainbows (*RGB rainbow*, *jet*, and *turbo*, in that order). These findings suggest that the more complex, interpretive tasks may depend on the cognitive characteristics of colormaps, rather than their perceptual appearance.

While our results conflict with conventional guidelines [51], the contrarian findings can largely be explained by our name variation metric. Rainbow colormaps, for instance, cross a broad range of readily identifiable colors (e.g., blue, red, orange, and yellow). This blend creates the appearance of "bands" that are easily distinguishable [39]. Although this kind of color discretization is believed to be problematic for quantitative data [5], we speculate that the visual system can take advantage of these emerging discrete features. For example, a viewer can heuristically use the apparent size, numeracy, and distribution of color patches as a proxy for various statistical properties in the data (e.g., variance [42]). People can reliably estimate these statistics with a quick glance likely due to fast-acting ensemble vision processes [1, 47, 55]. Such visual statistics may provide key summary features that help people discriminate between models. However, some ensemble processes operate solely on segmented visual features [16], which may explain the advantage for discretizing colormaps like rainbow. Our color name variation metric appears to model this tendency. Conversely, the smoother colormaps may complicate ensemble vision and potentially leave out key model-discriminating features. In addition to facilitating ensemble-based heuristics, color categorization may help observers encode specific data features in their working visual or verbal memories [24]. These mental representations could in turn be used to compare visualizations, enabling a more thorough assessment of different models. Future work is needed to test these visual ensembles and working memory explanations.

Name variation also explains performance discrepancies that we observed post-hoc. For instance, the difference between *viridis* and *plasma* is not well-explained by conventional design features. The two colormaps have virtually identical perceptual properties: both feature sequential, multi-hue, perceptually uniform ramps derived with spline interpolation in the CAM02 space [57]. Yet, results show *plasma* to be more accurate (68.6%, CI: 65.4–71.9) than *viridis* (66.4%, CI: 62–70.8). This performance gap cannot be explained with traditional design characteristics. However, the color name model readily distinguishes between these two ramps: it appears that *plasma* combines more nameable colors (3.44) than *viridis* (2.75). Intuitively, we observe shades or blue, purple, orange, and yellow in *plasma*, compared to a slightly more limited set of hues in *viridis* (navy blue, green, and yellow). Accordingly, and consistent with the observed data, the name model predicts slightly better inference accuracy with *plasma*.

The BIC criteria provides statistically robust evidence for a correlation between performance and name variation (§5.5.4). The comparison

against a conventional *design family* model suggests that traditional design properties, such as whether a colormap is sequential, divergent, or a 'rainbow' type, may not provide an optimal way of selecting colormaps. Graphical inference appears less dependent on these design characteristics. Instead, name variation alone may be sufficient to measure colormap utility for graphical inference.

7.2 Color Categorization & Specificity

Results from Experiment 1 indicate that color categorization may reveal subtle variations in the data. However, categorization could mislead viewers and cause them to misinterpret color boundaries as if they were data features [5] (**H2**). This misinterpretation could cause an analyst to falsely declare differences among visualizations that are otherwise sampled from the same model or phenomenon. Alternatively, an analyst could falsely conclude a discrepancy between a presumed prior model and a model represented by an observed visualization [17]. Yet, contrary to this hypothesis, Experiment 2 found no relation between name variation and specificity ($p = 0.464$). Participants were no more likely to report false differences while viewing colormaps with greater name variation. We find no evidence that the non-uniform colormaps (e.g., *RGB rainbow*) make people more susceptible to false positives compared to perceptually uniform ramps (e.g., *blues*, *viridis*, *coolwarm*). However, we did find a positive association between name variation and sensitivity (i.e., true positive rate), consistent with Experiment 1.

Findings from the two experiments could explain why rainbow and other hue-varying colormaps remain in wide use despite known limitations [35]. Our results suggest that the popularity of these schemes may be driven by greater utility for graphical inference as opposed to purely aesthetic preferences. This observation leads to an unorthodox guideline: designers may want to maximize the range of nameable colors in a ramp. The LAB-based model suggests an even simpler heuristic: maximizing a colormap's log-transformed curve length. While reasoning in log-LAB instead of LAB may appear counterintuitive, our analysis indicates that the former metric closely approximates name variation (Pearson's $r = 0.95$), providing a link between color appearance and name distance [20].

Rainbow schemes are not the only design family with broad name variation. Certain multi-hue colormaps, such as *plasma*, appear to strike a balance between perceptual uniformity and name variation. Inferences with *plasma* had above-average accuracy (68.6%), albeit slightly below *RGB rainbow* (70.3%, CI: 66.6–74.0). Accordingly, designers may opt for the more colorful maps while continuing to balance other design constraints, such as luminance monotonicity or perceptual uniformity. Though seemingly less relevant to graphical inference, these properties still play a key role in many tasks, including form perception [43,59]. Future work could investigate how to optimize these competing design constraints, for example, by computationally synthesizing new colormaps on demand (as in Colorgorical [18]). The two metrics proposed in this work (*color name variation* and *log-LAB length*) could aid this process by providing predictive models of colormap utility.

8 Limitations and Future Research

This study provides a first empirical investigation into how colormap design affects graphical inference. However, our approach provides only a preliminary lens onto this question and is subject to several limitations. We see these limitations as opportunities for future work.

Although we purposefully diversified the stimulus set by using a variety of distributions, all stimuli ultimately consisted of a mixture of Gaussians. These models approximate a range of phenomena in science and engineering, but they cannot capture all data types and characteristics. To further explore graphical inference in other contexts, the procedures employed in this study can be extended to tasks in specific domains. For example, to evaluate colormap effectiveness for flow visualization, an experiment could use flow models, applying measured perturbations to eddies and vortices to identify the most discriminating colormaps. For false-color astrophotography, an experiment could test models of galactic structures with perturbations to spirality or the homogeneity of galaxies.

Our study limits inference to a specific class of tasks known in the literature as "graphical inference" [10, 63]. Participants assessed ensemble visualizations holistically and made summary judgements. This task evaluates viewers' overall interpretations of a model and their ability to discriminate between different models. However, the task does not necessarily test if people can find subtle image differences. The term 'inference' can encompass a broad range of cognitive activities with visualizations (e.g., inductive reasoning and hypothesis testing), whereas we only consider a specific interpretation of this term.

Although our study shows a clear advantage for colorful ramps (e.g., rainbows), these ramps may not be accessible to people with color-vision deficiencies. Designers should therefore balance name variation with more accessible color metrics. One of the ramps we tested, *plasma*, seem to combine two accessible characteristics (relatively high name variation with monotonic luminance), making it a strong candidate for accessible visualization.

While results from Experiment 2 show that colorful maps do not necessarily decrease inference specificity, people could still perceive artifacts from rainbow maps in other contexts (e.g., when studying a single image, or when looking at multiple images with vastly different distributions). Understanding when such artifacts emerge and how they support or hinder analysis is key future work [39].

Lastly, our work considers *color name variation* as the main cognitive metric for modeling performance in graphical inference. However, we found that the results also correlate with log-LAB length. This latter metric, which may provide a measure of perceptual discriminative power, offers an alternative explanation for our observed results.

9 Conclusion

Color encoding is a fundamental concern for data visualization. Effective colormaps should not only help people perceive low-level data features, but must also facilitate accurate interpretation and inference. We theorized that performance in these tasks is correlated with the cognitive properties of colors. We tested this theory in two crowdsourced experiments measuring participants' ability to discriminate between 2D scalar models as a function of colormap design. We found that graphical inference is closely tied to the variety of nameable colors in the map. Colorful ramps, such as *jet* and *RGB rainbow*, led to significantly more accurate inferences than ramps comprising a limited selection of hues. In a second experiment, we measured the rate of false discovery but found no evidence of a relationship between color name variation and the false positive rate. Our results highlight the need for new color-encoding guidelines based on cognitive rather than purely perceptual factors. We proposed two new metrics for modeling colormap utility in visual inference. These metrics provide generative principles for new colormap designs while also extending conventional principles on color use in visualization.

10 Answers for Figure 1

Top row (starting with the leftmost lineup): top-right quadrant, bottom-left, bottom-right, top-left.
Bottom row: top-left, bottom-left, bottom-right, bottom-right.

Acknowledgements

This paper is based upon research supported by National Science Foundation awards 1942429, 1764092, & 1657599. The work was also supported in part by the Argonne Leadership Computing Facility, a U.S. Department of Energy Office of Science User Facility operated under contract DE-AC02-06CH11357.

References

[1] D. Ariely. Seeing sets: Representation by statistical properties. *Psychological Science*, 12(2):157–162, 2001.

[2] L. D. Bergman, B. E. Rogowitz, and L. A. Treinish. A rule-based tool for assisting colormap selection. In *Proceedings of the 6th conference on Visualization'95*, p. 118. IEEE Computer Society, 1995.

[3] K. P. Berlin B. Basic color terms: Their universality and evolution.-berkeley, 2002.

[4] M. Borkin, K. Gajos, A. Peters, D. Mitsouras, S. Melchionna, F. Rybicki, C. Feldman, and H. Pfister. Evaluation of artery visualizations for heart disease diagnosis. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2479–2488, 2011.

[5] D. Borland and R. M. T. Ii. Rainbow color map (still) considered harmful. *IEEE Computer Graphics and Applications*, 27(2), 2007.

[6] C. A. Brewer. Guidelines for selecting colors for diverging schemes on maps. *The Cartographic Journal*, 33(2):79–86, 1996.

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

[8] C. A. Brewer, A. M. MacEachren, L. W. Pickle, and D. Herrmann. Mapping mortality: Evaluating color schemes for choropleth maps. *Annals of the Association of American Geographers*, 87(3):411–438, 1997.

[9] A. M. Brown, D. T. Lindsey, and K. M. Guckes. Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. *Journal of vision*, 11(12):2–2, 2011.

[10] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4361–4383, 2009.

[11] 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*, 2017.

[12] N. R. Chowdhury, D. Cook, H. Hofmann, and M. Majumder. Measuring lineup difficulty by matching distance metrics with subject choices in crowd-sourced data. *Journal of Computational and Graphical Statistics*, 27(1):132–145, 2018.

[13] A. Dasgupta, J. Poco, B. Rogowitz, K. Han, E. Bertini, and C. T. Silva. The effect of color scales on climate scientists' objective and subjective performance in spatial data analysis tasks. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[14] R. De Valois and K. De Valouis. *Spatial Vision*. Oxford University Press, 1990.

[15] G. Derefeldt, T. Swartling, U. Berggrund, and P. Bodrogi. Cognitive color. *Color Research & Application*, 29(1):7–19, 2004.

[16] S. L. Franconeri, D. K. Bemis, and G. A. Alvarez. Number estimation relies on a set of segmented objects. *Cognition*, 113(1):1–13, 2009.

[17] A. Gelman. A bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review*, 71(2):369–382, 2003.

[18] 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, 2016.

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

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

[21] J. R. Hershey and P. A. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4, pp. IV–317. IEEE, 2007.

[22] H. Hofmann, L. Follett, M. Majumder, and D. Cook. Graphical tests for power comparison of competing designs. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2441–2448, 2012.

[23] M. D. Hyslop. A comparison of spectral color and greyscale continuous-tone map perception. Master's thesis, Michigan State University, 2006.

[24] T. Ikeda and N. Osaka. How are colors memorized in working memory? a functional magnetic resonance imaging study. *Neuroreport*, 18(2):111–114, 2007.

[25] B. L. Jones, D. S. Nagin, and K. Roeder. A sas procedure based on mixture models for estimating developmental trajectories. *Sociological methods & research*, 29(3):374–393, 2001.

[26] A. D. Kalvin, B. E. Rogowitz, A. Pelah, and A. Cohen. Building perceptual color maps for visualizing interval data. In *Human Vision and Electronic Imaging V*, vol. 3959, pp. 323–336. International Society for Optics and Photonics, 2000.

[27] J. Kim, D. J. Bodony, and J. B. Freund. Adjoint-based control of loud events in a turbulent jet. *Journal of Fluid Mechanics*, 741:28–59, 2014.

[28] G. Kindlmann and C. Scheidegger. An algebraic process for visualization design. *IEEE transactions on visualization and computer graphics*, 20(12):2181–2190, 2014.

[29] E. Komatsu, K. Smith, J. Dunkley, C. Bennett, B. Gold, G. Hinshaw, N. Jarosik, D. Larson, M. Nolta, L. Page, et al. Seven-year wilkinson microwave anisotropy probe (wmap*) observations: cosmological interpretation. *The Astrophysical Journal Supplement Series*, 192(2):18, 2011.

[30] M. P. Kumler and R. E. Groop. Continuous-tone mapping of smooth surfaces. *Cartography and Geographic Information Systems*, 17(4):279–289, 1990.

[31] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer. Selecting semantically-resonant colors for data visualization. In *Computer Graphics Forum*, vol. 32, pp. 401–410. Wiley Online Library, 2013.

[32] Y. Liu and J. Heer. Somewhere over the rainbow: An empirical assessment of quantitative colormaps. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 598. ACM, 2018.

[33] A. Mikhailov. Turbo, an improved rainbow colormap for visualization. `https://ai.googleblog.com/2019/08/turbo-improved-rainbow-colormap-for.html`, 2019. [Online; accessed 26-April-2020].

[34] K. Moreland. Diverging color maps for scientific visualization. In *International Symposium on Visual Computing*, pp. 92–103. Springer, 2009.

[35] K. Moreland. Why we use bad color maps and what you can do about it. *Electronic Imaging*, 2016(16):1–6, 2016.

[36] T. Munzner. *Visualization analysis and design*. CRC Press, 2014.

[37] 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, 2017.

[38] L. W. Pickle. Usability testing of map designs. In *Proceedings of Symposium on the Interface of Computing Science and Statistics*, pp. 42–56, 2003.

[39] P. S. Quinan, L. Padilla, S. H. Creem-Regehr, and M. Meyer. Examining implicit discretization in spectral schemes. In *Computer Graphics Forum*, vol. 38, pp. 363–374. Wiley Online Library, 2019.

[40] R. M. Ratwani, J. G. Trafton, and D. A. Boehm-Davis. Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied*, 14(1):36, 2008.

[41] K. Reda, P. Nalawade, and K. Ansah-Koi. Graphical perception of continuous quantitative maps: the effects of spatial frequency and colormap design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 272. ACM, 2018.

[42] K. Reda and M. E. Papka. Evaluating gradient perception in color-coded scalar fields. In *2019 IEEE Visualization Conference (VIS)*, pp. 271–275. IEEE, 2019.

[43] B. E. Rogowitz and A. D. Kalvin. The "which blair project": a quick visual method for evaluating perceptual color maps. In *Visualization, 2001. VIS'01. Proceedings*, pp. 183–556. IEEE, 2001.

[44] B. E. Rogowitz and L. A. Treinish. Using perceptual rules in interactive visualization. In *IS&T/SPIE 1994 International Symposium on Electronic Imaging: Science and Technology*, pp. 287–295. International Society for Optics and Photonics, 1994.

[45] B. E. Rogowitz and L. A. Treinish. Data visualization: the end of the rainbow. *IEEE spectrum*, 35(12):52–59, 1998.

[46] B. E. Rogowitz, L. A. Treinish, S. Bryson, et al. How not to lie with visualization. *Computers in Physics*, 10(3):268–273, 1996.

[47] J. Ross and D. C. Burr. Vision senses number directly. *Journal of vision*, 10(2):10–10, 2010.

[48] K. B. Schloss and S. E. Palmer. Aesthetic response to color combinations: preference, harmony, and similarity. *Attention, Perception, & Psychophysics*, 73(2):551–571, 2011.

[49] 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, 2015.

[50] H. M. Shapiro. *Practical flow cytometry*. John Wiley & Sons, 2005.

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

[52] S. Smart and D. A. Szafir. Measuring the separability of shape, size, and color in scatterplots. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019.

[53] 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.

[54] D. A. Szafir. Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[55] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri. Four types of ensemble coding in data visualizations. *Journal of vision*, 16(5):11–11, 2016.

[56] C. Tominski, G. Fuchs, and H. Schumann. Task-driven color coding. In *Information Visualisation, 2008. IV'08. 12th International Conference*, pp. 373–380. IEEE, 2008.

[57] S. van der Walt and N. Smith. Matplotlib colormaps. `https://bids.github.io/colormap/`, 2015. [Online; accessed 20-April-2020].

[58] W3C. Css values and units module level 3. `http://www.w3.org/TR/css3-values/#absolute-lengths`, 2016. [Online; accessed 20-April-2020].

[59] C. Ware. Color sequences for univariate maps: Theory, experiments and principles. *IEEE Computer Graphics and Applications*, 8(5):41–49, 1988.

[60] C. Ware. *Visual thinking: For design*. Morgan Kaufmann, 2010.

[61] C. Ware. *Information Visualization: Perception for Design*. Elsevier, 2012.

[62] C. Ware, T. L. Turton, R. Bujack, F. Samsel, P. Shrivastava, and D. H. Rogers. Measuring and modeling the feature detection threshold functions of colormaps. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[63] H. Wickham, D. Cook, H. Hofmann, and A. Buja. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):973–979, 2010.

[64] H. M. Widen, J. B. Elsner, S. Pau, and C. K. Uejio. Graphical inference in geographical research. *Geographical Analysis*, 48(2):115–131, 2016.

[65] M. Wijffelaars, R. Vliegen, J. J. Van Wijk, and E.-J. Van Der Linden. Generating color palettes using intuitive parameters. In *Computer Graphics Forum*, vol. 27, pp. 743–750. Wiley Online Library, 2008.

[66] D. N. Williams. Visualization and analysis tools for ultrascale climate data. *Eos, Transactions American Geophysical Union*, 95(42):377–378, 2014.

[67] J. Winawer, N. Witthoft, M. C. Frank, L. Wu, A. R. Wade, and L. Boroditsky. Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences*, 104(19):7780–7785, 2007.

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