Rainbow Colormaps: What are they *good* and *bad* for?

Khairi Reda

Abstract—Guidelines for color use in quantitative visualizations have strongly discouraged the use of rainbow colormaps, arguing instead for smooth designs that do not induce visual discontinuities or implicit color categories. However, the empirical evidence behind this argument has been mixed and, at times, even contradictory. In practice, rainbow colormaps are widely used, raising questions about the true utility or dangers of such designs. We study how color categorization impacts the interpretation of scalar fields. We first introduce an approach to detect latent categories in colormaps. We hypothesize that the appearance of color categories in scalar visualizations can be beneficial in that they enhance the perception of certain features, although at the cost of rendering other features less noticeable. In three crowdsourced experiments, we show that observers are more likely to discriminate global, distributional features when viewing colorful scales that induce categorization (e.g., rainbow or diverging schemes). Conversely, when seeing the same data through a less colorful representation, observers are more likely to report localized features defined by small variations in the data. Participants showed awareness of these different affordances and exhibited bias for exploiting the more discriminating colormap, given a particular feature type. Our results demonstrate the costs and benefits of rainbows (and similarly colorful schemes), suggesting that their complementary utility for analyzing scalar data should not be dismissed. In addition to explaining potentially valid uses of rainbow, our study provides actionable guidelines, including when such designs can be more harmful than useful. Data and materials are available at **`https://osf.io/xjhtf`**

Index Terms—Quantitative color encoding, rainbow colormaps, scalar fields, perception.

1 Introduction

Color mapping facilitates the visual encoding of data, allowing the latter to be represented with different gradations or categories of color. The question of how to design effective colormaps is a recurring theme in visualization research [55] and in other communities (e.g., oceanography, climate science, and astronomy [15, 22, 44]). Standard guidance indicates that quantitative colormaps should be perceptually uniform, orderable, and visually contiguous with no apparent color boundaries [9, 39]. Given these principles, the literature strongly discourages the use of rainbow colormaps [26], citing their 'non-scientific' basis [10] and their potential for introducing artifacts [5].

Much of this criticism has centered on the (non)perceptual properties of rainbows. In particular, their tendency to implicitly discrete data that should, ostensibly, be seen continuous [30]. The empirical evidence for or against rainbow colormaps, however, is inconsistent. Studies with human subjects have shown that rainbows can be advantageous in certain contexts [33], while others have turned evidence that they lead to inaccurate interpretation [24]. A possible explanation for these conflicting results is variation in the task: colormap evaluation studies have utilized a variety of tasks, from medical diagnosis [3], to estimation of summary statistics like the mean and spatial variance [11, 32], to low-level color discriminability [41, 43]. It is perhaps unsurprising that colormaps will vary in their usefulness depending on the task [45] or application domain [4]. Yet, a key limitation of extant studies is that it is often not possible to explain *why* a certain colormap design might have led to better (or worse) performance in a specific context. The lack of explanatory power makes it difficult to draw generalizable conclusions, especially to new tasks or domains that are yet to be studied.

An alternative to testing colormaps in different tasks is to specifically look for tradeoffs in colormap design and ask: what useful data features might be accentuated by a particular colormap design? We would expect different colormaps to make different types of features more (or less) visually prominent. We might also expect that in the process of highlighting certain features in the data, a colormap — whether rainbow or a more perceptually uniform design — can diminish the signal from other 'competing' features. Uncovering such tradeoffs will allow for a more nuanced colormap selection, whereby tasks are matched with suitable designs based on the feature(s) necessary for the task. Moreover, an empirical cost-utility account for color designs could potentially allow us to assimilate the conflicting evidence in colormap evaluation, where studies often reach contradictory conclusions (e.g., rainbows being both the worst [3, 24] and best-performing colormap [34]).

In this work, we investigate how one property of colormaps, *color categorization tendency*, affords varying interpretations of the same scalar data. Categorization occurs when a continuous color gradient is perceived (to some extent) as a sequence of discrete categories (e.g., blue, green, yellow). We evaluate the impact of such implicit discretization on the interpretation of two types of features in scalar fields: patterns characterized by *localized* variations in the density, and *global* distributional features, which dictate the overall spatial density of scalar data. We measured participants' sensitivity to these two features as we independently

• *Khairi Reda is with Indiana University–Purdue University Indianapolis. E-mail: redak@iu.edu*

manipulated them. We find that color categorization significantly influences the type of features participants attend to. In colormaps that induce categorization, participants exhibited a bias for detecting changes in the global structure of fields. This tendency was strongest in the rainbow, but the effect was also substantial in diverging and multi-hue colormaps. Conversely, scales that incorporate a few colors (e.g., single-hue schemes) afforded equal opportunity for detecting both localized and global features, albeit for an overall lower sensitivity. These results suggest that color categories imprint useful 'artifacts' in continuous fields, making it easier for observers to interpret structural properties with higher fidelity. However, these benefits come at the cost of reduced sensitivity to features defined by small, localized changes. We find that this effect is not limited to rainbows, but also manifests (to a lesser extent) in other perceptually uniform schemes, including diverging and multi-hue colormaps.

Our findings suggest that rainbow (and similarly colorful designs) should not be dismissed as inherently deceptive. Colorful scales appear to serve different functions, allowing observers to attend to different features in data, as compared with other designs traditionally deemed optimal. We confirm this hypothesis in two additional experiments and show that participants benefit substantially from seeing the same data encoded in two color schemes of complementary utility. More generally, we argue that a cost-benefit colormap analysis provides new insights that cannot be deduced from traditional color advice. We also introduce a methodology that can experimentally surface such tradeoffs. In sum, we make the following contributions:

1) We identify benefits and costs to implicit categorization in quantitative colormaps. Specifically, we show that color categorization increases the sensitivity to distributional features in data, at the expense of locally delineated features.
2) We introduce a new metric for quantifying color categorization. We show that this metric reliably predicts the type of features observers will attend to in scalar fields, irrespective of other colormap properties such as perceptual uniformity or luminance monotonicity.
3) We show that observers become more discerning when given the ability to interactively switch between two colormaps, suggesting a simple intervention for harnessing the benefits of different color encodings.
4) As a methodological contribution, we contribute experimental and data generation procedures to assess the cost-utility profiles of alternative colormap designs. These methods can be easily adapted to study other features or application domains of interest.

2 Related Work

We review guidelines for color encoding in visualization and discuss the results of extant colormap evaluation studies.

2.1 Color Advice for Visualization

Color design guidelines for visualizations have traditionally emphasized three properties of 'good' colormaps [9, 39]: *perceptual uniformity* (i.e., equal perceptual steps between adjacent colors), *continuity* (i.e., smooth color gradients), and *perceptual order* (i.e., intuitive color sequence, such as from dark red to bright yellow). A main reason why experts discourage the use of rainbow for interval data is that the colors are not perceptually orderable [38]. For instance, one cannot intuitively say whether green comes before or after blue, without explicitly consulting a legend. The lack of order makes it difficult to compare the magnitude in, say, two map points [14, 24]. On the other hand, the rainbow shines in value retrieval tasks [31, 50], making it easier for one to estimate the value associated with a specific color. Rainbows may also facilitate data recall compared to sequential schemes [14].

The desire for uniform yet colorful designs has led some to suggest 'spiral' colormaps: ramps that increase monotonically in luminance while rotating in the color wheel so as to incorporate multiple hues [51]. Compared to rainbows, multi-hue ramps are less colorful, although they typically retain perceptual uniformity. Luminance monotonicity also ensures the colors are intuitively orderable (e.g., from dark to bright). There has been a movement to end the use of rainbow [37], and replace it with alternative perceptual designs (e.g., *viridis* [46]). Researchers have also argued for diverging schemes, given the greater variation in luminance [7, 25]. Despite these efforts, rainbow colormaps continue to be widely used in practice [23, 26]. One explanation for this seemingly non-optimal choice is that people find rainbows to be aesthetically appealing. However, an alternative hypothesis that we explore here is that, despite their limitations, colorful encodings may provide some unrecognized utility.

Another common critique of rainbow designs is that they induce a form of implicit discretization [30]. This phenomenon, sometimes described as a hue banding effect, causes artificial boundaries between colors (e.g., between green and yellow). Color bands are thought to be undesirable because they are representationally incompatible with the underlying data (i.e., continuous data vs. seemingly discrete colors). More importantly, color boundaries could hypothetically cause people to infer false features that are not present in the data [5]. It is important to note, however, that implicit discretization is not an exclusive rainbow property. Most colormap designs, including perceptually uniform scales, will induce categorization to a certain extent (see Figure 2 for an illustration). It is unclear what benefits or perils such discretization may bring about — an issue that we explore in this work.

2.2 Empirical Colormap Evaluation

Empirical colormap studies have returned a variety of results that are, at times, contradictory. For example, Borkin et al. found rainbows to be inferior for interpreting arterial scan data [3]. Instead, doctors were more accurate at diagnosing heart disease with one of Brewer's diverging schemes, which are known for their perceptual uniformity [16]. Dasgupta et al. also show rainbows to be

less effective for estimating the mean magnitude in scalar fields [11]. On the other hand, Reda and Papka show rainbows to be effective for judging spatial variance [32]. Most surprisingly, standard rainbow designs (e.g., *jet* and *RGB rainbow*) were the best performing schemes in a task that required participants to discriminate models underlying a lineup of scalar fields [33,34]. It is difficult to reconcile these inconsistent results. One may attribute this variability to differences in the task; it is generally accepted that different tasks will benefit from different color encodings [4,45]. Still, the above results do not tell us when and why a certain task might benefit from, say, a perceptually uniform scheme versus a more colorful encoding. Standard color guidelines are also not tailored to tasks, but instead simply consider the general data type (e.g., interval, ratio, or categorical) [10,27]. To reconcile these differences, we consider the features a task might require. We argue that different colormaps will accentuate different aspects of the data. This in turn makes them more or less suited for a given task, depending on how well a colormap conveys data features required by a task.

2.3 Feature-Driven Colormap Evaluation

A number of studies have attempted to assess feature discriminability in colormaps. Rogowitz et al. show that detecting Gaussians requires a consistent level of luminance-contrast against the background [36]. They accordingly suggest that quantitative colormaps should increase monotonically in luminance [35]. Kalvin et al. tested the perceptibility of Gabor features in different colormaps, this time modulating the feature's spatial frequency. They argue that the luminance channel is best suited for representing high-spatial frequency features, while saturation is more effective for conveying low-spatial frequency information [1, 21]. In more recent work, Ware et al. found that feature detection threshold is a function of local perceptual differences (including both in chroma and luminance), although their model overweighs luminance [52].

This work similarly considers the discriminability of features in scalar fields. However, in addition to testing how a class of features might be enhanced by certain colormap designs, we consider other feature types that might be degraded in the process. Compared to earlier experiments on feature discriminability [21,52], our task is also different in that participants do not know the location of features in advance. Instead, they are forced to compare and classify the pattern in multiple plots in order to identify one of two concealed targets. This 'lineup' task is thought to require a combination of visuo-spatial and cognitive reasoning skills [47]. In addition to small, localized features [52], our setup also includes larger distributional features that compete for participants' attention. This combination is meant to simulate a more realistic scenario where there are both low-frequency (i.e., global structure) as well as high-frequency (localized) information. Participants in our study had to weigh these competing signals before making a judgment. We explain the results using a metric of color categorization. As tasks become more interpretive, cognitive color properties like names and categories [12] are likely to take prominence.

2.4 Graphical Inference for Colormap Evaluation

Empirical studies have utilized graphical inference [8,19] as a model interpretive task for evaluating colormaps. Most similar to our work are studies by Reda and Szafir who found colormaps that cross a variety of color names to be more effective [34]. A second study looked at the effects of color name salience, finding better performance for color scales that incorporate more salient names (e.g., prototypical 'yellow' over 'beige') [33]. While these studies demonstrate advantages to colorful encodings (including rainbows), the results are one-sided in that they only show benefits, but without documenting potential side-effects. These earlier results are also difficult to explain and reconcile with standard color advice (e.g., 'rainbow colormaps considered harmful' [5,37]). We address this gap by introducing a modified inference task that includes multiple targets shown concurrently. This new stimulus format allows us to characterize potentially different affordances for colormaps, rather than measuring performance along a single dimension. Specifically, we extend earlier work [34] in three important ways. First, we simultaneously study two classes of features (global versus localized). We show that color categorization yields benefits but also exerts costs, in the form of enhanced perceptibility for certain data features at the cost of others. In effect, we contribute evidence that there are not inherently 'good' or 'bad' colormaps, but only instead that different colormaps will highlight different things in data. This framework for colormap evaluation, which admits both costs and benefits to alternative designs, helps explain the conflicting empirical results from earlier studies (e.g., rainbows documented as both the worst [3] and best [34] design). Second, we contribute a metric for quantifying implicit categorization in continuous colormaps, leveraging empirical color-naming models [18]. We show that this metric reliably explains participants' bias to one class of features over another, across a variety of colormaps. Third, we evaluate a simple intervention for leveraging complementary color designs, by allowing observers to interactively switch the colormap. We show that even inexperienced observers can effectively choose the most discerning colormap for a given dataset. In sum, we contribute a more nuanced theory and evidence for how to choose effective colors for scalar fields.

3 Research Questions and Methods

We address two research questions in this work:

RQ 1: Given alternative color mapping strategies (e.g., single-hue versus more colorful rainbow schemes), are there identifiable costs and benefits to these designs? In this paper, we are especially interested in color categorization: the tendency for an observer to see discrete colors in a gradient. We ask whether such implicit categorization can aid or hinder the interpretation of continuous (scalar) data.

Humans intuitively "parse the continuum of color into discrete categories" [40]. While quantitative colormaps are designed to appear as smooth gradients, most viewers will perceive color categories within, especially in the more colorful scales. Rainbow colormaps are mostly known for inducing categorization, which presents as a sequence of

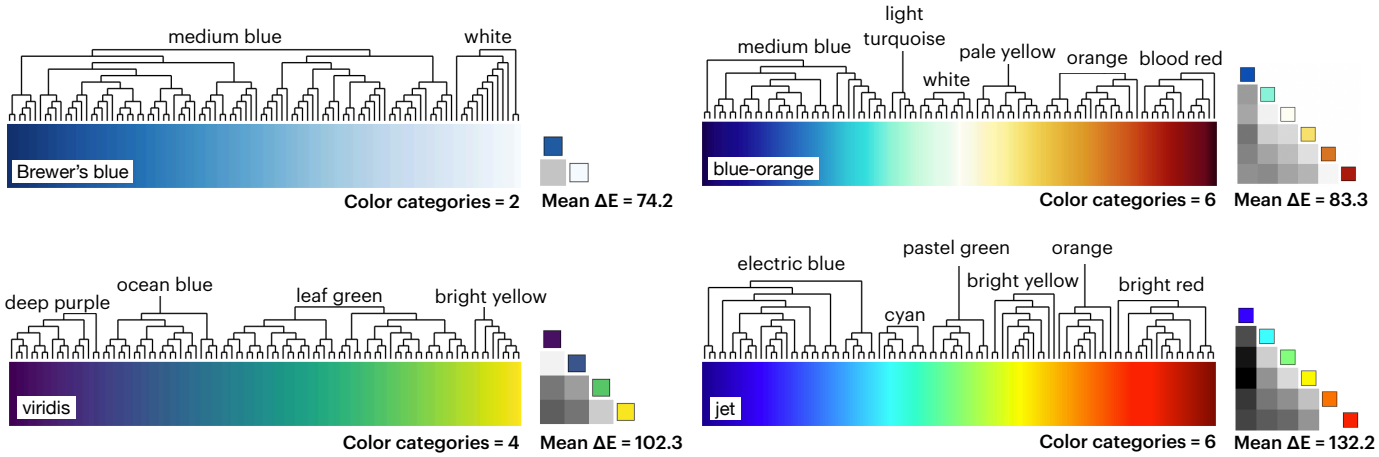

Fig. 1. Illustration of our color categorization metric applied to four continuous colormaps. Dendrograms show hierarchical clustering results. Top-level clusters represent the final color categories detected. Color names shown (extracted from the Heer and Stone model [18]) represent the most likely names for the centroids of top-level clusters. Matrices show pairwise perceptual distances (ΔE) between the centroids (darker is higher discriminability). We define **color categorization tendency** as the **number of color categories** times their **mean ΔE distance**. For instance, *jet* has a categorization tendency of 6 $\times$ 132.2 = 793.2.

blue, green, yellow, orange, and red. However, discretization also occurs, to a lesser extent, in other colormap designs. For instance, a diverging *blue-orange* scale displays distinct tones of blue, cyan, white, orange, and dark red. On the other hand, single-hue colormaps like Brewer's *blue* exhibit markedly less categorization. As an example of a tradeoff in colormap design, we ask if color categorization can enhance the perception of certain data features, perhaps at the expense of others.

RQ2: If implicit color categories can indeed enhance the perception of certain features in data while suppressing others, can observers detect this modulating effect? This question has practical importance: if observers are aware of the costs and benefits to different colormaps, they might benefit from seeing two representations of the same data. On a theoretical level, an observer who effectively leverages multiple color encodings provides compelling evidence of different affordances for alternative colormap designs, even in one task and for the same general data type.

We address the first question in Experiments 1 and 2, where we expose participants to stimuli containing two competing features. We measure participants' potential bias to one feature type over the other. In Experiment 3, we add the ability for participants to interactively switch between two colormaps (representation varying levels of categorization). We test the effect of this addition on participants' accuracy and colormap selection behavior. Before we set out to describe the three experiments, we discuss our approach to measuring color categorization. We then describe the overall experimental apparatus and discuss our hypotheses.

3.1 Modeling Color Categorization

Color categorization is thought to originate in the early layers of the visual cortex [42], suggesting a possible biological basis [40]. However, evidence for color categories is arguably easier to ascertain in the languages we speak [29, 54]. Across numerous languages, prominent colors (e.g., red, green, yellow) are associated with unique and stable names [2, 17]. We thus quantify the level of implicit categorization for a colormap by first clustering unique color names that appear within. This gives us an approximation of the number of categories a viewer might perceive. We subsequently measure the perceptual distance between those categories to estimate how discriminable they are. Compared to earlier approaches for quantifying categorization (e.g., locally summing up color names or perceptual distances [34]), this new metric provides two advantages: first, it represents a principled approach of initially detecting and grouping latent categories and then measuring their global (as opposed to local) discriminability. Global discriminability in turn allows estimating whether those categories can be uniquely attended to and addressed by an observer. Second, while this approach for quantifying categorization tendency is more computationally intensive, it provides for more easily explainable results than a low-level metric alone, such as *log-LAB Length* [34]. We illustrate how this new metric works with four colormaps (Figure 1).

3.1.1 Detecting Latent Color Categories

To detect latent color categories in a colormap, we first sample the color scale at evenly spaced key points, obtaining a sequence of n discrete colors. We choose $n = 100$ for our analysis. We then cluster neighboring colors according to their name: adjacent colors with similar names are progressively merged into larger clusters. This method is equivalent to bottom-up (agglomerative) hierarchical clustering. We use a greedy implementation, merging two colors (or color clusters) that have the highest name similarity at every step, and stopping when there are no similar clusters to merge. We quantify name similarity using Heer and Stone's cosine name distance metric [18]. This metric returns the empirical probability that two given colors are associated with distinct names. For example, $\Delta Name(\blacksquare,\blacksquare) = 0.17$, meaning there is only 17% chance that these two colors carry different names, making them a candidate for merge. By contrast $\Delta Name(\blacksquare,\blacksquare) = 0.64$, so most people will likely see them as two distinct categories (e.g., blue and turquoise). We stop the merge when name dissimilarity for a color relative to its

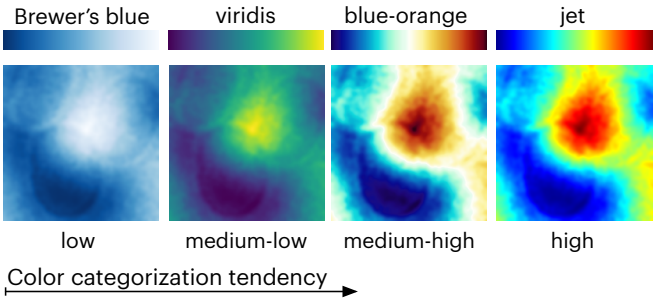

Fig. 2. Four colormaps (used in Exp. 1) representing increasing levels of color categorization. Note the effect of categorization on the perceptibility of features in the scalar fields: the 'bean'-shaped features are most visible in Brewer's *blue*, but those features become less discernible as color categorization increases. However, categorization may afford a better overview of the data's spatial extent and distribution.

neighbors exceeds a certain threshold, which we manually set to 0.6. That is, two adjacent colors (or clusters) will merge unless the probability they have different names is slightly higher than chance.

For each emerging cluster, we designate the member color with the most *salient name* as the centroid, with the latter used to measure cluster-cluster name distance. Name saliency is the degree to which a color is associated with a uniquely recognizable name. More prototypical colors (e.g., 'red' and 'yellow') have higher name saliency over other similar tones (e.g., 'crimson' and 'beige'). We use Heer and Stone's saliency metric [18] which maps name saliency for CIELAB colors to a [0, 1] range (1 being most salient). To illustrate, a cluster consisting of (■,■,■) will have ■ designated as centroid, given that the latter is associated with an empirically more salient name. The rationale behind adopting salient names for centroids is to encourage clustering around readily nameable colors. We found that name saliency yields more representative centroids compared to the geometric CIELAB mean color for a cluster — the latter tends to dilute centroids towards neutral chromas.

We interpret the final (top-level) clusters as representatives of the latent color categories for a colormap. Figure 1 illustrates the emerging categories for four color scales. Our method appears to detect color bands seen in typical designs. For example, in *jet*, we find six categories representing blue, cyan, green, yellow, orange, and red (see Figure 1). The categories detected also match up with bands appearing in scalar fields (see Figure 2). For instance, in *blue-orange*, we observe color bands representing blue, turquoise, white, yellow, orange, and dark red, which correspond to categories detected by our clustering procedure. Note that our procedure correctly classifies sharp color boundaries (e.g., the white 'mach' band in *blue-orange*) as well as smoother gradients combining distinctively nameable tones (e.g., the transition from cyan to green in *jet*). As expected, we detect fewer categories in *viridis* and Brewer's *blue* (four and two categories, respectively). The diverging *blue-orange* appears to have the same number of categories (six) as *jet*, although the latter has more distinctive categories. For instance, the color difference between bright yellow and orange in *jet* is seemingly higher relative to the difference between pale yellow and orange in *blue-orange*. To account for this variation,

TABLE 1
Characteristics of four colormaps used in Experiment 1, including color categorization tendency (rightmost column). The latter is defined as the number of color categories times their mean ΔE category distance.

Colormap	Design	Color Categories	Mean ΔE	Categorization Tendency
Brewer's blue	Single-hue (monotonic luminance)	2	74.2	148.4
viridis	Multi-hue (monotonic luminance)	4	102.3	409.2
blue-orange	Diverging	6	83.3	499.8
jet	Rainbow	6	132.2	793.2

we measure the perceptual distance between the resultant categories and operationalize ΔE in our metric.

3.1.2 Quantifying Color Categorization Tendency

Although the number of nameable colors provides important information to approximate categorization tendency, those categories must still be discriminable by an observer to play a role in analysis. Therefore, after clustering, we compute the perceptual distance between the resultant category centroids. Specifically, we compute the CIELAB ΔE distance between every pair of top-level clusters detected in the earlier step. Figure 1 illustrates the pairwise category distance as matrices. Summing each half-matrix and dividing by the number of cells ($\frac{K(K-1)}{2}$; with K as the number of categories detected), we obtain a measure of the mean, relative discriminability for the categories. To illustrate, the mean ΔE distance for *jet* is 132.2. This is markedly higher than *blue-orange* (83.3), *virids* (102.2) and Brewer's *blue* (74.2).

Lastly, we operationalize both of the above components into a single metric of *color categorization tendency*. Specifically, we weigh the number of categories by their mean perceptual discriminability:

$$\textbf{Color Categorization Tendency} = K \times \text{Mean Category Discriminability} = \frac{2}{(K-1)} \sum_{i=2}^{K} \sum_{j=1}^{i} \Delta E(c_i, c_j)$$

Where K is the number of color categories (i.e., top-level clusters) detected, c_i is the centroid color for category i. We use CIE76 ΔE (Euclidean distance in the LAB space). By combining the above two components (K and *mean* ΔE), we obtain a measure of how many discriminable categories there are in a colormap. In particular, operationalizing perceptual discriminability allows us to gauge how easy is it for a viewer to attend to each color category separately — a key hypothesized driver behind the benefits for implicit categorization (see §4). We later show that this combined metric explains the experimental results (see §5, §6).

Table 1 gives the color categorization tendencies for the four colormaps. As expected, *jet* ranks highest in its degree of implicit categorization, followed by *blue-orange* and *viridis*. Among the four designs, Brewer's *blue* exhibits the least categorization. Note also that categorization is not necessarily reduced by adopting a perceptually uniform gradient. For example, both *viridis* and *blue-orange* are uniform (equal perceptual steps). Yet, according to our metric,

viridis can be classified as a medium-low whereas *blue-orange* exhibits medium-high categorization.

3.2 Experimental Task

To evaluate the impact of implicit colormap categorization, we devised a modified 'lineup' task based on Buja et al's concept of graphical inference [8, 53]. In the original task, the observer compares a 'real' dataset to data that had been sampled from a null model. This is achieved by concealing a visualization of the real dataset in a lineup that also includes a number of 'decoy' plots representing null samples. The observer is then challenged to identify the plot that looks different (i.e., the visualization of the real data). An observer who correctly identifies the real target provides evidence that the data at hand is statistically different from the null. In effect, lineups are a visual analog to null hypothesis significance testing.

Lineup tasks have been used to evaluate the power of different colormap designs. Colormaps that make it easier to discriminate the target plot (without increasing the rate of false positives) were deemed more effective [34]. However, earlier lineup experiments did not allow for studying tradeoffs in colormap design. This is because the target and decoy plots differed in one aspect: the spatial distribution of scalar data. We extend that work by modeling sensitivity to features defined by localized variations. The latter, in particular, represents an important class of features relevant to scientific applications.

To experimentally model the tension between localized patterns versus global structure, we adopt a modified lineup format that incorporates two targets (i.e., two different-looking visualizations as opposed to one) [48]. The first target in our setup represents a visualization that varies systematically from the decoys in the global distribution of its scalar data. We induce the latter by varying the layout of density-generating processes. The second target represents a visualization where localized features vary from those that are present in the decoy (and in target 1, for that matter). Specifically, we change the orientation of processes that embed low-amplitude patterns in the density. The latter presents as oval-shaped features whose outline is defined by small perturbations in the global density (see Figures 2 and 3).

Note that, in this setup, both features are present in all plots. However, to an observer, a lineup stimulus conveys two competing signals: a target that is distinctively different in its global density (hereafter, the 'global target') and a second that is unique in its configuration of localized patterns (i.e., a 'local target'). By presenting both targets alongside decoys, we can detect any potential bias in attending to localized over global structures (or vice versa). For example, if a particular color design makes it more likely for participants to report the global target, we would conclude an increased tendency for discriminating distributional features with that encoding. Conversely, were participants to show a propensity to report the local target, we can infer a bias for discriminating localized features. We describe how we generated lineup stimuli to incorporate the two features. To ensure both features are equally discriminable, we calibrated their saliency in a pre-study (see §5.3).

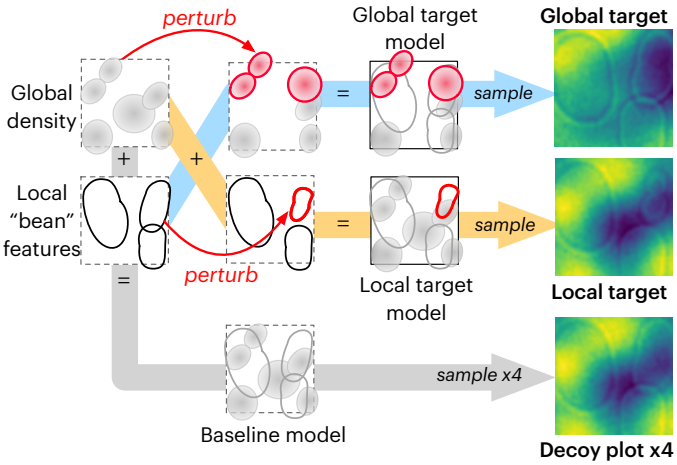

Fig. 3. Overview of the stimulus synthesis procedure. Global density and localized 'bean' shaped features are first synthesized as two separate layers, and integrated into a baseline model (grey arrow). The feature layers are then independently perturbed to produce a *global target* model (incorporating *perturbed global* + baseline local features; blue arrow) and an opponent *local target* model (with baseline global + *perturbed local* features). The models are finally sampled to generate a lineup with four decoys and two targets. Note that the magnitude of perturbation was exaggerated in this figure for illustration; actual perturbation in our experiments was less obvious in both targets.

3.3 Feature Design & Stimulus Synthesis

To embed known global and local structures in scalar fields, we designed two data-generating processes. The first process dictates the overall structure (i.e., spatial distribution). The second process introduces localized perturbations in the above distribution, in effect imprinting features outlined by low-amplitude variations in the global density.

3.3.1 Global Density Features

We first synthesize a global density model using a mixture of Gaussians: clusters of 2D Gaussian kernels are positioned randomly in the 2D domain. Kernel size (i.e., standard deviation) and orientation (i.e., X-Y correlation for a kernel) are varied randomly within a set range (determined through piloting). The kernels are randomly assigned to generate either 'positive' or 'negative' densities, which are then accumulated to precipitate the global structure for a scalar field.

3.3.2 Localized Features

We synthesize a second 'layer' containing a different set of features whose outline is marked by localized variation in the density. We similarly start with randomly configured clusters of Gaussian kernels. However, instead of accumulating the density from those kernels, we use them as scaffolds to generate 2D sine-wave impulses. We center each impulse around a Gaussian cluster, rendering it at a randomly chosen contour level. This process creates semi-irregular oval-shaped features (which were referred to by some participants as 'beans'; see Figure 2).

Compared to Gabor and Gaussian features used in colormap evaluation [21, 52], our bean-shaped pattern is similarly defined by a high-frequency, low-contrast impulse. However, our chosen features are wider in spatial extent, causing them to span a larger part of the colormap. This design allows us to model a situation where the observer needs

to comprehend a complex pattern while also accounting for the potential effects of color categorization. The beans are also reminiscent of a 'ripple' pattern sometimes used for colormap testing [28], although we purposely designed our features to be non-symmetric (pilot tests showed that symmetric shapes were far easier to discern, which would have limited our ability to control their relative salience).

3.3.3 Lineup Generation

The global and localized feature layers are added together into one scalar field. The field is treated as a 2D probability distribution model from which random samples are drawn to generate plots for the lineup. Importantly, the two data-generating processes described above allow us to manipulate the two feature types independently. Starting from a baseline model (comprising both global and localized features), we randomly translate the location of the density-generating kernels (by a set amount, calibrated empirically in §5.3), resulting in a model that differs in its global spatial distribution relative to the baseline. We sample this perturbed model to generate a first target plot. We similarly perturb the localized-feature process, this time adjusting the contour level at which the sine impulses are rendered. This causes the bean-shaped features to change in extent and/or orientation (see Figure 4-left for an illustration of this effect). We sample from this perturbed model to obtain a second target. Lastly, we sample the baseline model four times to obtain four decoys. The process is illustrated in Figure 3.

The two targets are randomly placed alongside the decoys in one lineup (for a total of six plots per lineup). Participants are prompted to select the visualization that "doesn't belong". Even though there were two targets in each lineup, we limited participants to a single selection for simplicity. Earlier work shows that observers rarely select more than one answer in a dual-target lineup task [48]. The process above produces a large variety of stimuli owing to the stochastic generation and interactions between Gaussian mixtures. Note also that, in addition to the principal variation between the targets and the decoys, the decoy plots also vary among themselves due to random sampling. These sources of variation make for a complex classification task that recruits both cognitive and visuo-spatial skills [47].

4 Hypotheses

We developed two hypotheses:

H1 — Viewing a colormap with high categorization tendency (e.g., *jet* or *blue-orange*), participants will identify the global target more frequently. The latter represents variations in the global density relative to a baseline model. Conversely, when viewing a low-categorization colormap (e.g., *blue* or *viridis*), participants will select the local target more often. In other words, we expect the ratio of global to local targets reported to correlate with color categorization tendency (as defined in §3.1).

Color categories are technically 'artifacts' that are not actually present in the data. However, they represent a useful kind of artifact in that they make it easier to attend to specific data subsets. For example, a viewer could focus on peaks by selectively attending to red. Similarly, one could attend to the middle or low values of the distribution, if those are associated with separable colors. Indeed, theories of visual attention suggest a utility for color categories here. The boolean map theory, for instance, stipulates that observers can only process multi-color displays serially, each time attending to a spatial region defined by a single color [20]. In effect, spatially distributed features can only be accessed one color at a time. Accordingly, an observer may find it easier to assess the spatial structure in a scalar field, if the display exhibits clearly segmentable colors. By contrast, visualizations exhibiting smooth color gradients may not readily support the creation of boolean color masks, hence affording less attention to distributional features.

Although useful for highlighting structural properties, color categorization could obscure other types of features. Specifically, categorical distinction might overwhelm smaller, within-category color differences. In such displays, we might expect degraded discriminability for features or patterns defined by localized, graduated perturbations in the global structure. In particular, we would expect the 'bean' patterns to become less noticeable with *jet* or *blue-orange*. By contrast, low-categorization schemes (e.g., Brewer's *blue* and *viridis*) do not suffer from this issue as much, and should thus make it relatively easier to discriminate the local target.

H2 — We expect participants to show awareness of the above tradeoffs. Specifically, we envision an alternative lineup setup where only one type of variation occurs (i.e., either global or localized discriminating features are present). If participants are given the ability to choose the colormap, we would expect them to select the most discerning color scheme for a particular stimulus. This hypothesis suggests a simple intervention to overcoming the limitations of a single color encoding: rather than providing one representation that might not adequately convey all relevant features, designers should redundantly encode the same data in multiple colormaps (or, where possible, allow the viewer to interactively toggle between colormaps). The hypothesis also reinforces the idea that different colormaps afford varying interpretations of the data, even when the task and data characteristics are largely identical.

We test H1 in Exp. 1 and 2. We subsequently adapt the experimental apparatus, adding the ability to interactively switch between colormaps, and test H2 in Exp. 3.

5 Experiment I

We evaluated the impact of color categorization on the interpretation of scalar fields in a crowdsourced experiment. We adopt the lineup task described in §3.2, incorporating both global and localized features in every trial. We conjecture that color categorization would help participants resolve subtle differences in the global structure of scalar fields. However, we also conjecture that this advantage would come at the cost of decreased sensitivity to features defined by small, localized changes.

5.1 Experiment Design

We employ a within-subject design: participants saw four different colormap designs representing different levels of color categorization (from low to high): Brewer's *blue*,

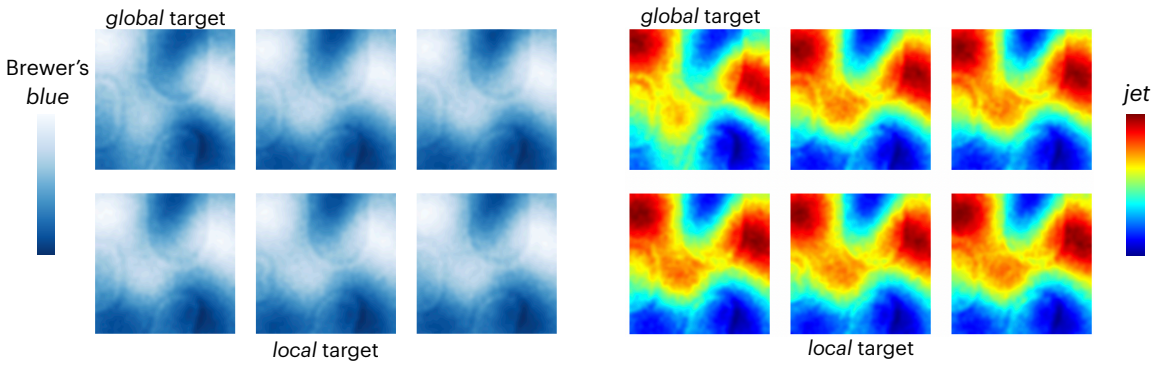

Fig. 4. Example stimuli from Experiment 1 shown using low- and high-categorization colormaps. A stimulus comprised a lineup of six scalar fields, two of which (the *targets*) vary systematically from the rest although in different ways. Our experiments demonstrate bias among participants for selecting one type of target over the other, depending on the colormap. For instance, participants overwhelmingly select the 'global' target with *jet* whereas they are equally likely to identify local or global targets with Brewer's *blue*, albeit at lower sensitivity. In effect, we show that different colormaps afford varying yet equally valuable interpretations of the same data.

viridis, *blue-orange*, and *jet* (see Figure 2). The experiment was blocked by colormap, with block order varied randomly. Participants completed 9 trials with each block, for a total of 36 trials. A trial comprised a freshly generated lineup of six scalar fields (200×200 pixels each), presented in a 3×2 arrangement. All lineups consisted of four decoys and two targets, representing global and localized variations from the decoys (see Figure 4). A color scale was displayed to the right of the lineup for reference. The relative saliency for the two targets was equalized and subsequently maintained in all trials (see §5.3 for details). Participants were instructed to select "the image that doesn't belong". They made their choice by clicking one of the six plots and confirmed their selection by pressing Enter, before moving to the next trial.

5.2 Procedure

Participants first completed a color-vision qualification test (14 Ishihara panels), followed by a brief tutorial. They then completed 12 practice trials each containing a single target: a field that was varied either in its global or localized structure relative to the decoys. We limited practice trials to one target at a time in order to cue participants about the two target types individually. During practice, participants were given feedback, informing them if they had guessed correctly or otherwise pointing out the correct answer. After practice, participants completed the analyzed trials in which no feedback was provided. To ensure participants were paying attention, we included 2 engagement checks per block (8 checks in total). The checks consisted of single-target lineups where the scalar density in one of the visualizations was inverted (i.e., a 'negative' image of the decoys), making for an easy judgment. After completing four blocks of analyzed trials, participants were asked to briefly describe their strategy. Lastly, participants filled out a demographic survey. We excluded participants who failed most of the engagement checks and recruited new participants in replacement, until we reached our intended sample size (see §5.4).

5.3 Calibrating Feature Salience

To ensure equal baseline saliency for the two targets, we conducted a pair of pre-studies in which we independently calibrated the discriminability of the two feature types. In the first pre-study, we varied the perturbation for global density features. This was done by randomly offsetting the center of density-generating Gaussians in the target while keeping local (bean) features unchanged. This process results in only a single 'valid' target that differs in its global density. We varied the level of perturbation in the target across trials from low to large. Higher perturbation implies a target that is more dissimilar from the decoys, making for easier judgment. We used the *viridis* colormap for all pre-study trials, owing to its uniformity and status as a default colormap in many visualization systems.

We recruited 50 participants from Amazon Mechanical Turk, dropping 3 due to incomplete data. We fitted the results using logistic regression, modeling the likelihood of target discrimination in response to perturbation magnitude. By offsetting kernel centers with a random vector of $length$=9%$\times min(W_{field}, H_{field})$, we achieved an expected accuracy of 58% (halfway between $\frac{1}{6}$ chance and perfect reliability). We then conducted a second pre-study with 50 participants (2 of whom were dropped), this time varying the perturbation for the localized 'bean' features while keeping the global density unchanged. A logistic regression model indicated that perturbing one of three beans led to an expected accuracy of approximately 58%, similar to the detection threshold achieved earlier for global features. We subsequently use these parameters for the main experiment, generating two targets in every lineup according to the above perturbation levels.

5.4 Participants

For the main experiment, we recruited 70 participants from Amazon Mechanical Turk (41 males, 28 females, and 1 did not identify). The average reported age for participants was 39.4 years ($\sigma = 10.9$). We limited enrollment to US residents with a task-approval history of at least 97%. Participants received a compensation of $3. We dropped 2 participants because their data was incomplete, leaving 68 in the analysis.

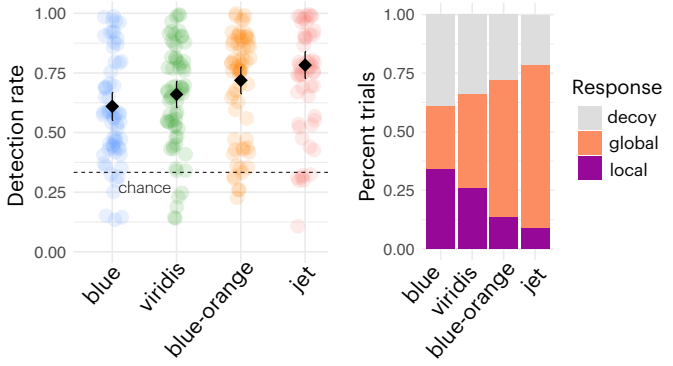

Fig. 5. **Left**: Detection rate for correctly identifying *either* of the two targets by colormap. Circles show rates for individual participants. Diamonds depict group averages (± 95% confidence intervals). **Right**: Percent trials in which participants reported the local target (purple), global target (red) and, incorrectly, the decoy (grey). Colormaps are ordered from left to right based on categorization tendency.

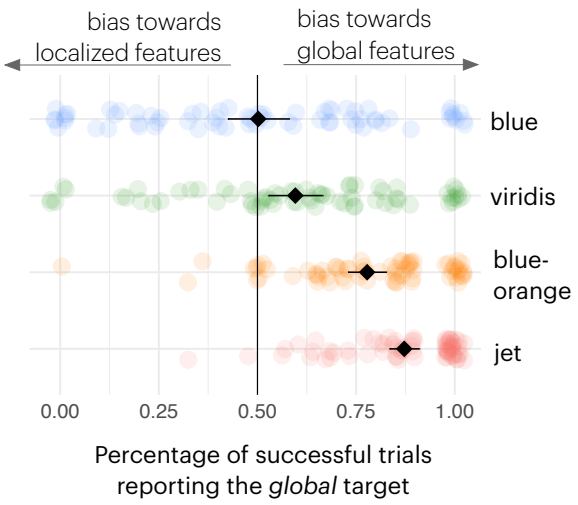

Fig. 6. Bias for reporting global over local targets by colormap (ordered from top to bottom based on their categorization tendency). Circles show the bias for individual participants. Diamonds depict group means (± 95% CI). A percentage of over 50% indicates bias for global features.

5.5 Results

Participants completed the experiment in 19 minutes on average ($\sigma = 10.2$). They collectively provided 2,448 responses. We first analyze the overall rate of detection and subsequently look for evidence of bias for either target type in response to color categorization.

5.5.1 Detection Rate

Participants successfully detected one of the two targets in 69.28% of trials ($\sigma = 18.58\%$). Figure 5–left plots detection rate by colormap. We modeled the results using a logistic regression model, with *color categorization tendency* as a fixed effect (see Table 1). We apply a log-transform to the latter to improve model fit. We also include two random intercepts in the model capturing individual differences and variation due to repeated trials. The model indicates a significant effect of color categorization score (Wald's $Z = 7.158$, $p < 0.001$). A unit increase in the latter increases the odds of correctly identifying either target by 1.71 (95% CI: 1.47—1.99). Overall, participants performed better when viewing lineups with a colorful ramp (e.g., *jet* or *blue-orange* over Brewer's *blue* and *viridis*).

5.5.2 Global vs. Localized Features

On average, participants identified the global target in 47.41% of trials ($\sigma = 16.86\%$). By contrast, the local target was reported in only 20.55% of trials ($\sigma = 12.20\%$). The remainder ($31.65\%, \sigma = 18.2\%$) represent unsuccessful trials where participants incorrectly selected one of the decoys. Beyond the average, there was a substantial difference in the ratio of global to local targets reported across the four colormaps (see Figure 5–right).

We model the likelihood of choosing *global* over *local* targets as a function of color categorization. Figure 6 illustrates this relationship. We include responses in which either target was identified, and build a logistic regression model with *color categorization tendency* as a fixed effect (log-transformed). We also include two random intercepts to account for individual differences and trial order (to isolate any residual learning effects). The model indicates a significant effect for color categorization ($Z = 13.747$, $p < 0.001$). A unit increase in log-categorization tendency (equivalent to adding approximately 3 distinct color names to the scale) causes a 3.82-fold increase (95% confidence intervals: 3.16—4.63) in the odds of discriminating global *over* local features. The above tendency can be clearly seen in individual colormaps. For example, in a successful trial with Brewer's *blue*, the average participant had a 50.2% (CI: 42.5—58%) chance of reporting the global target. In other words, it was equally likely for participants to identify localized and global discriminating features when viewing a low-categorization colormap. By contrast, the global target was reported in 87.2% (CI: 83.5—91%) of successful trials with *jet*. Similarly, there was substantial bias for the global target with *blue-orange* (77.8% of successful trials, CI: 73—82.6%). The two latter colormaps exhibit higher categorization. While also significant, the bias with *viridis* was less pronounced (59.6%, CI: 52.8—66.5%).

5.5.3 Discussion

Overall, participants were more likely to discriminate global as opposed to localized features. This is possibly due to the latter requiring more effort and acuity overall. More interestingly, however, we see a marked difference in the makeup of targets reported in the four colormaps. Specifically, there is a tendency for participants to over-report the global target when looking at a colorful visualization. This tendency was clearest in rainbow, although we still see the same pattern in *blue-orange*. That the latter exhibits similar bias is significant, and suggests that, irrespective of perceptual uniformity, color categorization helps to highlight distributional features in scalar fields. By contrast, Brewer's *blue* exhibited a balanced response profile, with similar numbers of local and global targets reported. This suggests equal salience for both feature types. The results support H1, although the discriminability of localized features with *blue* and *viridis* was actually less than we had anticipated. Accordingly, overall sensitivity was lower in the less colorful designs.

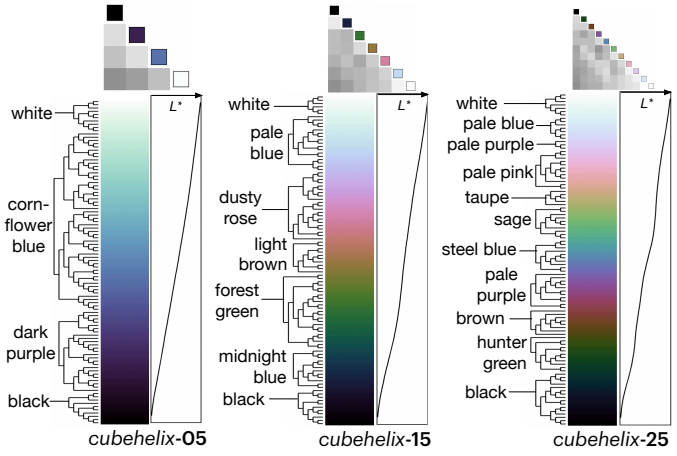

Fig. 7. Three *cubehelix* variants used in Exp. 2. The scales comprise 0.5, 1.5, and 2.5 hue rotations, giving rise to increasing levels of color categorization (measured at *253.4*, *435.2*, and *629.6*, respectively). However, all three colormaps have approximately linear luminance ramps (L^* profile shown to the right of each color scale).

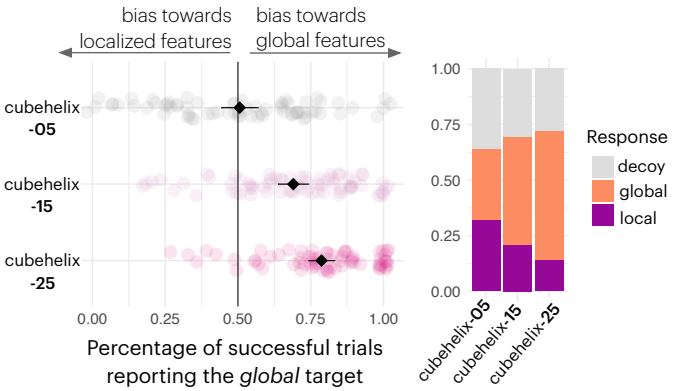

Fig. 8. **Left:** Bias for reporting the global over the local target in Exp. 2. Colormaps are ordered based on their categorization tendency (top to bottom). Circles depict bias for individual participants. Diamonds are group means (± 95% CI). **Right:** Distribution of responses (global, local target, or decoy) by colormap.

One limitation of this experiment is the diverse set of colormap designs tested, which could present a confound. We address this issue in a follow-up experiment.

6 Experiment II

Exp. 1 shows that color categorization can affect the type of features observers attend to. Yet, this result may have be confounded because the colormaps tested varied not just in their level of categorization, but also in their design and luminance profile. For instance, Brewer's *blue* and *viridis* comprise linear luminance ramps. By contrast, *blue-orange* is a diverging scale with peak luminance in the middle. Luminance in *jet* varies irregularly. To rule out this potential confound, we conducted a second experiment where we tested colormaps of similar design, but that which vary systematically in their color categorization tendency. Specifically, we utilize variants of *cubehelix* [15], a multi-hue color scale that parameterizes the number of hue rotations. By varying the latter, we obtain colormaps with increasing levels of categorization but without affecting luminance. We generate three *cubehelix* scales with 0.5, 1.5, and 2.5 hue rotations (1.5 rotations yields the default *cubehelix* scale [15]). As the number of hue rotations increases, so does color categorization. The three colormaps, however, share approximately the same monotonic luminance profile (see Figure 7). We attempt to replicate the results of Exp. 1 with these three colormaps. We adopt the same lineup task, incorporating two targets in every trial. We expect the level of categorization to dictate the type of targets reported by participants. In particular *cubehelix-25* should lead to greater reporting of global targets (i.e., similar to *jet* and *blue-orange*). By contrast, we would expect *cubehelix-05* to yield equal numbers of global and local targets (i.e., much like *blue*).

6.1 Participants, Experiment Design, and Procedures

We recruited 70 participants from Amazon Mechanical Turk (43 females, 27 males; mean age of 35 years). We limited enrollment to US residents who have at least 97% task-approval rate. Participants had to pass a color-vision qualification test. They were compensated with a $3 payment. The experiment was a within-subject design: all participants saw the three *cubehelix* variants presented in 3 blocks. Block order was randomized. Each block consisted of 12 trials, for a total of 36 trials, plus 9 engagement checks distributed evenly between the blocks. Similar to Exp. 1, each trial comprised a lineup of six scalar fields with two targets (corresponding to global and localized features) and four decoys. Participants were instructed to select the "image that doesn't belong." Participants first completed 12 practice trials with feedback, followed by 36 analyzed trials.

6.2 Results

Participants completed the experiment in 20.18 minutes on average ($\sigma = 9.92$). The global target was reported in 46.27% of trials compared to 22.18% for the local target. However, similar to Exp. 1, the distribution of targets reported varied substantially across the three colormaps (see Figure 8). We analyzed the results using a logistic regression model, with (log-transformed) color categorization tendency as a predictor. The model indicates a significant effect for color categorization (Wald's $Z = 11.217, p < 0.001$). A step in the latter increases the odds of reporting global targets by a factor of 5.7 (95% CI: 4.2—7.72). Accordingly, the majority of targets (78.7%) reported in *cubehelix-25* (high categorization) were global. By contrast, the ratio of global to local target in *cubehelix-05* (low categorization) was roughly 1:1. *Cubehelix-15* was in the middle with 69% global targets.

6.3 Discussion

The above results replicate the effects observed in Exp. 1: color categorization is associated with a bias towards global, distributional features. A colormap's categorization tendency is predictive of the degree of bias. These results provide additional evidence to support H1. Moreover, because all *cubehelix* colormaps are derived from the same underlying design (i.e., multi-hue with approximately monotonic luminance), we can rule out a potential confound in

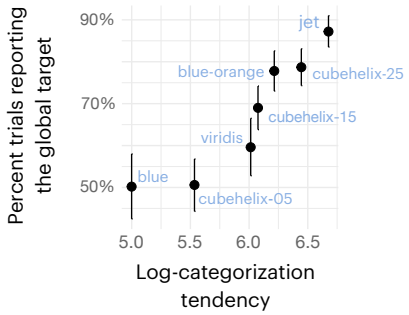

Fig. 9. Bias towards global features as a function of color categorization. Circles depict colormaps in Experiments 1 and 2. Error bars depict 95% CI for the mean observed bias.

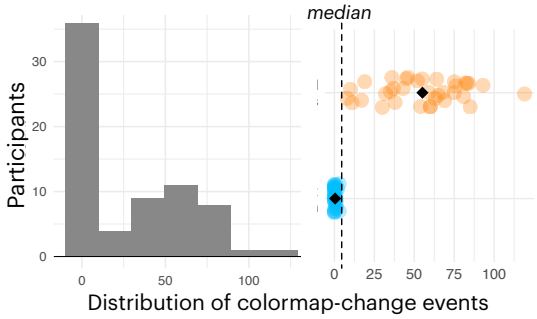

Fig. 10. A histogram of the frequency of colormap switching shows a bimodal distribution. Half the participants did not frequently switch the colormap during the experiment (blue dots), with the other half having actively utilized this interaction (orange). We separate participants into these two groups ($n = 35$ each) along the median of the distribution.

Exp. 1. In fact, color categorization exerted very similar effects across the two experiments. For instance, the performance profile of *cubehelix-25* is virtually identical to *blue-orange* (global target reported in 78.7% and 77.8% of trials, respectively), despite differences in colormap design. The similarity in performance, however, can be explained by the (log-transformed) categorization scores for the two colormaps, which are comparable (6.44 and 6.21, respectively). Lastly, the estimated effect size for categorization in the two experiments overlaps (odds ratio 3.16—4.63 in Exp. 1 versus 4.2—7.72 in Exp. 2). Figure 9 plots the combined results (seven colormaps) from the two experiments. Color categorization appears to reliably predict the degree of bias.

7 Experiment III

Experiments 1 and 2 provide evidence of different affordances for colormaps. Colorful designs that induce categorization allow better resolution of distributional features in scalar fields. Less colorful scales, on the other hand, appear to afford equal attention to both global and localized patterns, albeit for a generally lower sensitivity. It is unclear, however, if observers can be made aware of these tradeoffs, or whether they can actively leverage two color designs in the same task. In this experiment, we simulate a context where people could benefit from multiple color encodings. We employ the same lineup task (§3.2), but instead of two targets in every lineup, we embed a single target that differs either in its overall density or in its localized pattern configuration. We allow participants to interactively toggle between two colormaps: Brewer's *blue* and *blue-orange*, which exhibited complementary roles in Exp. 1. We test if participants can take advantage of this dual-colormap setup, and whether they show willingness to exploit the more discriminating colormap for a given target.

7.1 Participants, Experiment Design, and Procedures

We recruited 70 participants from Amazon Mechanical Turk (41 males, 27 females, and 2 did not identify; mean age of 37.4 years). We limited enrollment to US residents with a task-approval rate of at least 97%. We compensated participants with a $3 payment. The experiment comprised 4 blocks of 8 trials each, for a total of 32 trials. In half of the trials, we embedded a single target that differed from the decoys in its spatial distribution. The other half contained a target varying in the configuration of its bean-shaped features. The two target types appeared equally within each block. Trial order was randomized. At the beginning of each trial, we randomly set the initial colormap to either *blue* or *blue-orange*. Participants were then allowed to interactively switch the colormap by pressing any key on their keyboard. This caused the six visualizations in the lineup to be redrawn instantaneously using the new colormap. Participants were prompted to select the "image that doesn't belong." The prompt also reminded participants that they could switch the colormap if they wished. There was no limit on how many times a participant could cycle between the two colormaps in a given trial. In addition to recording correctness for each trial (i.e., whether the target was correctly identified), we tracked three metrics: the amount of time a participant spends viewing each colormap, the 'final' colormap with which the participant made their target selection, and the number of colormap switches per trial.

Participants first completed a color-visio test, followed by a tutorial and a 12-trial practice session. After practice, they completed the 32 analyzed trials (plus 8 engagement checks) and finished the experiment by describing their strategy and filling out a short demographic survey.

7.2 Results

Participants completed the experiment in 22.5 minutes on average ($\sigma = 12.2$), with an overall accuracy of 50.98%. We look at how frequently participants switched the colormap, and analyze the effects of this behavior on target detection.

7.2.1 Colormap Switching Behavior

The frequency of colormap switching shows a bimodal distribution in which half the participants did not change the colormap or did so only a handful of times (see Figure 10). The other half appear to have frequently changed the colormap (as high as 119 times during the experiment). We split participants along the median (5.5 switches) into two equally sized groups: those who switch the colormap ($\mu = 55.1$ switches, $\sigma = 26.1$) and those who rarely do, if at all ($\mu = 0.457$ switches, $\sigma = 0.817$). The latter group relied mostly on the randomly assigned colormap in each

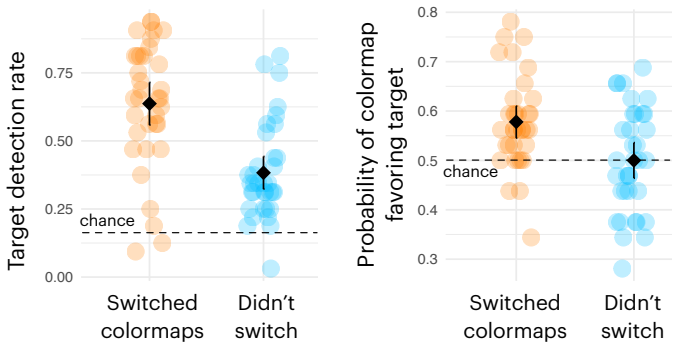

Fig. 11. **Left**: difference in target detection rate between participants who frequently switch colormaps (orange) and those who do not (blue). **Right**: the probability a participant settles on a 'final' colormap favoring the particular target they had been observing. Each circle represents one participant. Diamonds depict group means (± 95% CIs).

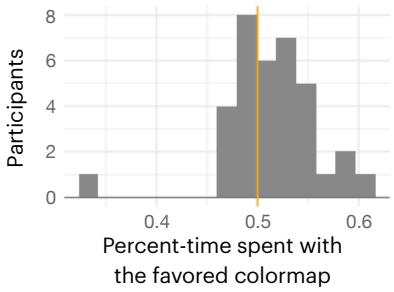

Fig. 12. The fraction of time participants looked at stimuli with a colormap favoring the target. Deviation from 50% indicates a tendency for exploiting the more discriminating colormap for a given target.

trail. We model the likelihood of successful target discrimination based on the above switching behavior. Using logistic regression, we find that colormap switching is associated with a higher likelihood of successful discrimination (Wald's $Z = 5.048$, $p < 0.001$). Specifically, participants who changed the colormap improved the odds of identifying the correct target by a factor of 3.28 (95% CI: 2.07—5.2), leading to markedly higher accuracy ($\mu = 63.7\%$, CI: 55.8—71.5%) compared to those who mostly settled for the default colormap ($\mu = 38.3\%$, CI: 32.3—44.3%). Figure 11-left illustrates the relationship between successful detection and colormap switching. To rule out a confound owing to different levels of engagement, we compared the accuracy of the two groups on the engagement checks. Participants who actively switched the colormap had a slightly higher engagement rate (96.4% vs. 91.1%). The difference, however, was small ($p > 0.05$), amounting to a mere 0.4 missed attention-test on average. Participants who changed the colormap had significantly longer response times (23.3 vs. 16.4 seconds on average, $p < 0.001$). However, this behavior is to be expected and suggests active use of two encodings.

7.2.2 Bias for Favored Colormaps

We looked for evidence of bias towards the more effective colormap. Recall that the two colormaps available to participants favor different target types: while we expect *blue-orange* to be useful in resolving globally distinct targets, Brewer's *blue* should favor targets that vary in their localized feature configuration. We anticipate that, after switching, participants will settle on the most discriminating color scheme in a specific trial. We found that those who switch colormaps are in fact 1.37 (95% CI: 1.14—1.66, $Z = 3.28$, $p < 0.01$) times more likely to settle on a final scale that favors the lineup they had been observing at the time. By contrast, and as would be expected from a random assignment, we find no evidence of bias among those who rarely switched colors (see Figure 11-right).

Lastly, we analyzed the fraction of time participants utilized the favored colormap for a specific lineup. We limit this analysis to those who used the colormap switching feature. Figure 12 illustrates this distribution. In an average trial, participants utilized the favored colormap for 51.22% of their response time (95% CI: 49.61—52.83%). This slight bias, however, was not significantly different from an even 50-50 split ($t(34) = 1.535$, $p = 0.134$).

7.2.3 Discussion

The results provide further evidence of differing utilities for colormaps, depending on their color categorization tendency. While only half the participants showed willingness to switch colormaps, those who did appear to have gained a significant advantage at the task, amounting to more than a 3-fold improvement at finding the real target. These results provide further evidence to support H1. Participants who switched colors also exhibited bias for exploiting the more discriminating colormap for a particular target, suggesting an awareness of the complementary utilities for the two encodings. These results support H2. That said, it is not clear why half the participants did not take advantage of this feature (i.e., colormap switching). One possibility is that, despite our encouragement, some participants did not think that changing colors would help. Another possibility is that, because changing colors increases response time, many participants forgo this option (crowd-workers often have limited bandwidth to allocate for each task).

8 General Discussion

Color advice for visualization consists primarily of generic rules of thumb (e.g., no apparent color boundaries when representing continuous data). However, it is unclear if and why these rules should equally apply to different tasks. Empirical studies have so far painted a mixed picture of the utility of these principles. For instance, the often criticized rainbow can be both the best [34] and worst performing colormap [3], depending on the task. We currently lack a framework to parse this conflicting evidence, which cannot be reconciled under existing rules. In this work, we proposed a theoretical framework for colormap evaluation in which we seek to characterize both costs and benefits for alternative colormap designs. Specifically, we theorized that different colormaps can accentuate different kinds of features in data. However, by making certain features more prominent, a colormap can also diminish the signal from other competing features. We developed experimental procedures to test this framework, and conducted three experiments to characterize the effects of color categorization on the interpretation of scalar fields.

8.1 The Costs and Benefits of Colorful Designs

Color categorization is the tendency to see discrete colors in an otherwise continuous color gradient. Although this effect is often associated with rainbows, in fact, most colormaps will bring about some degree of implicit discretization. Traditionally, such categorization is considered only appropriate for categorical data, but recent work suggests broader utility. For instance, in graphical inference, people are better at discriminating distributions when those are visualized with rainbow [34]. However, along with these benefits, we hypothesize that there are costs to color categorization, in that the latter would make it harder to see patterns defined by localized variations. Exp. 1 and 2 illustrate this tradeoff: viewing a colorful scale (i.e., high categorization), participants were more likely to identify a global target that varies in its distribution. Simultaneously, those same participants were far less likely to notice features defined by small perturbations in the distribution.

The bias above lays bare the benefits and costs of color categorization. A colorful representation makes it possible to see differences in how scalar data is globally distributed. For instance, it is easier to check if there are subtle variations in how the peaks or valleys are distributed in multiple visualizations. Simultaneously, however, categorization seems to reduce one's ability to sense smaller changes in color. Consequently, features marked by small, relative differences in scalar data (e.g., the 'beans' in Figure 2) are particularly harder to discern. Outcompeted by larger color differences, the localized targets became less discernible in *jet* and *blue-orange*. Note importantly that, in the presence of two feature types, the colorful *blue-orange* induces a similar bias to global features as *jet*, despite the former's perceptually uniform profile. It is also notable that even a slight degree of categorization appears to bring about a proportional bias for global features, as in *viridis*. By contrast, the least colorful scales (Brewer's *blue* and *cubehelix-05*) showed no bias, instead allowing participants to equally discern global and localized features. However, overall sensitivity was significantly reduced in the latter scales. For all colormap designs tested, the above tendencies can be explained by a single metric of color categorization.

The results are consistent with H1, which predicted the aforementioned effect for color categorization. However, the bias towards localized features in *blue* was smaller than we had anticipated, presenting instead as equal saliency for both feature types. This resulted in uneven utility for the four colormaps in Exp. 1. A potential explanation is the local targets require more effort to identify, which skews the advantage in favor of the more colorful designs. That said, color categorization was strongly predictive of the degree of bias towards global features. Exp. 2 replicated this effect in three *cubehelix* color scales, which share the same design characteristics but vary in their categorization tendency. Exp. 2 therefore enables us to attribute the results to color categorization, as opposed to other colormap design factors. Taken together, the experiments lend support to H1.

One might have expected *blue-orange* (Exp. 1) to be superior for resolving localized patterns, given its high discriminative power. Ware et al. report *blue-orange* as one of the better colormaps for detecting high-frequency features [52]. A potential factor behind the lower detection rate here is the complexity of the 'bean' pattern. Compared to Ware et al's Gabors, the bean spanned a larger spatial extent, often crossing multiple color bands. There were also multiple beans in each plot, with only one presenting as uniquely different in the target. As such, identifying the correct target involved not only low-level feature detection but also attentional and short-term memory resources (e.g., for visual comparison). The presence of strong color boundaries could crowd out this feature, and potentially disrupt contour integration [13], hence complicating the discovery of localized targets. On the other hand, color categories give rise to a stronger Gestalt for distributional features, in effect enhancing the saliency of the global target. The competition between two feature types in this study uniquely illustrates a tradeoff that was not modeled in earlier work [21,34,52].

8.2 Rethinking Color Advice for Visualization

Our results back up some of the existing color advice for visualization. For example, the argument that rainbows can obscure small features [37] appears to have merit. Interestingly, we observed this tendency not just in rainbows, but also in other colorful designs. In fact, one could predict the bias against localized features by measuring color categorization tendency alone, irrespective of other colormap characteristics like perceptual uniformity or luminance monotonicity. Therefore, the cautionary advice above should extend to all schemes that cross multiple color categories, including diverging and multi-hue designs.

Although the discriminability of small features is essential in many domains, global features may be just as important. The latter appear to benefit substantially from color categorization. Here, *jet* outperformed all other designs for interpreting distributional characteristics. This may explain why rainbow colormaps continue to be used widely in practice, despite advice to the contrary [5,10,23,26]. Accordingly, in lieu of a dichotomy of 'good' versus 'bad' colormaps, we should recognize unique affordances for alternative colormap designs. Given a task or a communication goal, we can then recommend a specific color design based on what we know to be the relevant features. For example, in datasets containing important low-contrast features, we might recommend a monotonic luminance ramp that crosses very few, if any, color categories. On the other hand, if the goal is to broadly convey distributional features, we might recommend a more colorful scale spanning multiple color categories (e.g., *blue-orange* or even a rainbow design). We suspect that many practitioners already understand such tradeoffs at an intuitive level. Our study contributes empirical evidence of these tradeoffs along with a new predictive metric of color categorization.

Different colormaps appear to emphasize different things in data, potentially giving rise to varying (though equally valuable) interpretations. We further speculate that different color encodings allow an observer to build different mental models of the same data. A colorful representation favors creating a model that highlights distributional characteristics. Here, the location of features such as peaks, valleys, and middle values is emphasized in the model. By contrast, a less colorful representation leaves little impression of the spatial distribution, thus allowing low-amplitude

features to stand out. Accordingly, we might recommend redundant encoding where the same data is visualized twice using two different colormaps (e.g., a colorful *and* a smooth-looking scale with few color categories). This dual-colormap arrangement may serve to represent different aspects of the data, just like how multiple views are often used in visualization interfaces [49]. Exp. 3 tested an analogous intervention, allowing participants to toggle between two colormaps. Participants who actively switched the colormap benefited from this straightforward intervention. The advantage presented as a 3-fold increase in the likelihood of resolving the target. Furthermore, and consistent with H2, we obtained evidence that participants were aware of complementary utilities for the colormaps that had been available to them. This awareness manifested as a tendency to exploit the more discriminating colormap, given a particular target. Exp. 3 thus reinforces the idea of unique affordances for different colormaps, even when the task is unchanged.

8.3 Accessibility for Color Vision Deficiency

One important cost to colorful scales is the reduced accessibility for people with color-vision deficiency (CVD). Rainbow schemes can be particularly problematic here. It is thus essential to consider accessibility in the cost-benefit analysis, especially when designing graphics intended for a wide audience. In particular, it may be advisable to refrain from using rainbow colormaps exclusively for figures in the scientific literature [23], or in critical domains (e.g., communication in weather emergencies). Fortunately, our results suggest other CVD-friendly designs that share similar qualities to rainbow, in terms of effectively communicating distributional features. While not fully matching the color nameability of the latter, diverging and certain multi-hue schemes (e.g., *blue-orange* and *cubehelix-25*) appear to induce a useful degree of color categorization, and may thus be used in place of *jet*. If color nameability is essential, it may be appropriate to visualize the same data with two color scales one of which is a rainbow and the other a CVD-friendly scheme. Lastly, it may be possible to obtain CVD-friendly rainbow variants by truncating problematic color tones (e.g., yellow-greens) [6]. This allows for retaining a colorful design while ensuring accessibility for most viewers.

9 Limitations & Future Work

There are limitations to our work that should be contextualized. First, we opted to limit participants to a single target selection. This was meant to simplify the response structure. Similar dual-choice experiments suggest that participants will mostly self-limit their response to a single choice [48]. Because of this, Exp. 1 and 2 measure *relative salience* for the two features (i.e., global vs. localized), rather than modeling *absolute* feature discriminability. While saliency can predict what aspects an observer attends to, visual attention is also driven by other factors, including top-down processes. Accordingly, despite the bias to global features, localized features may still be sufficiently discriminable in *blue-orange* or even in *jet*. Our results do not exclude the latter possibility, but instead show attentional bias to one class of features over the other, subject to color categorization. Second, given the nature of the lineup task, our results are most applicable to situations where the observer is actively comparing multiple plots. Though we anticipate similar biases to manifest, it would be interesting to evaluate those tendencies in an alternative setup involving single visualizations. While we evaluate two classes of features for scalar fields, future work might attempt to replicate our results with other exemplifying features, or with entirely different feature classes. The effects of color categorization could also be studied in other data displays, including standard charts. To that end, the methods developed in this work should be relatively easy to adapt to other kinds of features or visualizations. Lastly, the metric we propose in this work is only meant to capture one aspect of quantitative colormaps (i.e., implicit categorization tendency). Although we show this property to be important, there are, too, other considerations for continuous colormaps. Our metric should therefore be considered alongside other properties (e.g., smoothness and local discriminative power), which may be equally important to consider depending on the task.

10 Conclusion

Color advice for visualization consists of rules that match colormap designs to generic data types (e.g., monotonically increasing luminance for quantitative data vs. hue-varying palettes for nominal). However, these rules do not consider task nuances, giving rise to inconsistent results in formal evaluations, or to the advice being ignored by practitioners. We proposed an alternative colormap evaluation framework in which we characterize both costs and benefits, thereby allowing for unique affordances to competing colormap designs, even in the same task. The results of three experiments confirm the usefulness of this approach. Specifically, we show that colorful scales (like rainbow and diverging schemes) can accentuate the distributional characteristics of scalar fields, while simultaneously rendering localized features less discriminable. Our work contributes experimental methods and metrics for assessing the cost-utility of colormaps. We also contribute to an ongoing debate on the appropriateness of rainbow colormaps for visualization.

Acknowledgments

This paper is based upon research supported by the National Science Foundation under award 1942429.

References

[1] 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.
[2] B. Berlin and P. Kay. *Basic color terms: Their universality and evolution*. Univ of California Press, 1991.
[3] 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.
[4] D. Borland and A. Huber. Collaboration-specific color-map design. *IEEE Computer Graphics and Applications*, 31(4):7–11, 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. Brewer. Spectral schemes: Controversial color use on maps. *Cartography and Geographic Information Systems*, 24(4):203–220, 1997.

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

[8] 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.

[9] 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.

[10] F. Crameri, G. E. Shephard, and P. J. Heron. The misuse of colour in science communication. *Nature communications*, 11(1):1–10, 2020.

[11] 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.

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

[13] D. J. Field, A. Hayes, and R. F. Hess. Contour integration by the human visual system: evidence for a local "association field". *Vision research*, 33(2):173–193, 1993.

[14] I. Golebiowska and A. Coltekin. Rainbow dash: Intuitiveness, interpretability and memorability of the rainbow color scheme in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[15] D. A. Green. A colour scheme for the display of astronomical intensity images. *Bulletin of the Astronomical Society of India*, 39:289–295, June 2011.

[16] 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.

[17] D. G. Hays, E. Margolis, R. Naroll, and D. R. Perkins. Color term salience. *American Anthropologist*, 74(5):1107–1121, 1972.

[18] 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.

[19] 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.

[20] L. Huang and H. Pashler. A boolean map theory of visual attention. *Psychological review*, 114(3):599, 2007.

[21] 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.

[22] N. Kaye, A. Hartley, and D. Hemming. Mapping the climate: guidance on appropriate techniques to map climate variables and their uncertainty. *Geoscientific Model Development*, 5(1):245–256, 2012.

[23] A. Light and P. J. Bartlein. The end of the rainbow? color schemes for improved data graphics. *Eos, Transactions American Geophysical Union*, 85(40):385–391, 2004.

[24] 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.

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

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

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

[28] P. Nardini, M. Chen, R. Bujack, M. Bottinger, and G. Scheuermann. A testing environment for continuous colormaps. *IEEE transactions on visualization and computer graphics*, 27(2):1043–1053, 2020.

[29] G. V. Paramei. Singing the russian blues: An argument for culturally basic color terms. *Cross-cultural research*, 39(1):10–38, 2005.

[30] 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.

[31] 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.

[32] 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.

[33] K. Reda, A. A. Salvi, J. Gray, and M. E. Papka. Color nameability predicts inference accuracy in spatial visualizations. In *Computer Graphics Forum*, vol. 40, pp. 49–60. Wiley Online Library, 2021.

[34] K. Reda and D. A. Szafir. Rainbows revisited: Modeling effective colormap design for graphical inference. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1032–1042, 2021. doi: 10.1109/TVCG.2020.3030439

[35] 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.

[36] B. E. Rogowitz, A. D. Kalvin, A. Pelah, and A. Cohen. Which trajectories through which perceptually uniform color spaces produce appropriate colors scales for interval data? In *Color and Imaging Conference*, vol. 1999, pp. 321–326. Society for Imaging Science and Technology, 1999.

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

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

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

[40] A. E. Skelton, G. Catchpole, J. T. Abbott, J. M. Bosten, and A. Franklin. Biological origins of color categorization. *Proceedings of the National Academy of Sciences*, 114(21):5545–5550, 2017.

[41] 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.

[42] M. Sun, L. Hu, X. Xin, and X. Zhang. Neural hierarchy of color categorization: From prototype encoding to boundary encoding. *Frontiers in neuroscience*, p. 861, 2021.

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

[44] K. M. Thyng, C. A. Greene, R. D. Hetland, H. M. Zimmerle, and S. F. DiMarco. True colors of oceanography: Guidelines for effective and accurate colormap selection. *Oceanography*, 29(3):9–13, 2016.

[45] 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.

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

[47] S. VanderPlas and H. Hofmann. Spatial reasoning and data displays. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):459–468, 2015.

[48] S. VanderPlas and H. Hofmann. Clusters beat trend!? testing feature hierarchy in statistical graphics. *Journal of Computational and Graphical Statistics*, 26(2):231–242, 2017.

[49] M. Q. Wang Baldonado, A. Woodruff, and A. Kuchinsky. Guidelines for using multiple views in information visualization. In *Proceedings of the working conference on Advanced visual interfaces*, pp. 110–119, 2000.

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

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

[52] 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.

[53] 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.

[54] 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.

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

Khairi Reda is an Associate Professor in the School of Informatics & Computing at Indiana University–Purdue University Indianapolis. His research lies at the intersection of Human-Computer Interaction and Data Science, with interests in visual analytics and data cognition.