Modeling Color Difference for Visualization Design

Danielle Albers Szafir, *Member, IEEE*

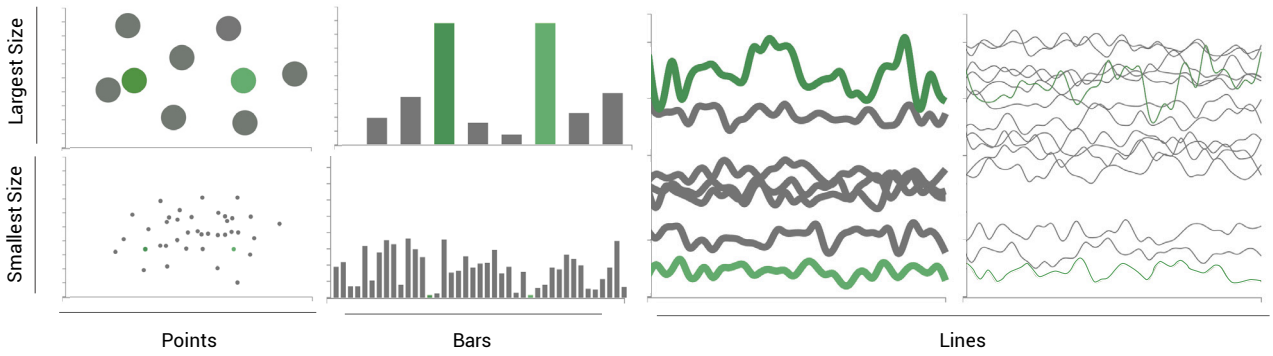

Fig. 1. We performed three experiments to measure color difference perceptions for visualizations, focusing on diagonally symmetric marks from scatterplots, elongated marks from bar charts, and asymmetric elongated marks from line graphs. The tested size ranges are shown above for two greens at $\Delta E = 10$ (figures have been scaled to 30% of the tested size). We confirm prior findings that perceived color difference varies inversely with size and find that elongated marks provide significantly greater discriminability for encoding designers. Our results provide probabilistic models of color difference for visualization.

Abstract—Color is frequently used to encode values in visualizations. For color encodings to be effective, the mapping between colors and values must preserve important differences in the data. However, most guidelines for effective color choice in visualization are based on either color perceptions measured using large, uniform fields in optimal viewing environments or on qualitative intuitions. These limitations may cause data misinterpretation in visualizations, which frequently use small, elongated marks. Our goal is to develop quantitative metrics to help people use color more effectively in visualizations. We present a series of crowdsourced studies measuring color difference perceptions for three common mark types: points, bars, and lines. Our results indicate that peoples' abilities to perceive color differences varies significantly across mark types. Probabilistic models constructed from the resulting data can provide objective guidance for designers, allowing them to anticipate viewer perceptions in order to inform effective encoding design.

Index Terms—Color Perception, Graphical Perception, Color Models, Crowdsourcing

1 Introduction

Visualizations reveal patterns in data by mapping values to different visual channels, such as position, size, or color. In order for visualizations to be effective, perceived differences in encoded values should correspond to differences in the underlying data. As a result, visualization designers need to map data ranges to sufficiently wide ranges in the target visual channel such that important differences in the data are preserved. However, most metrics for predicting perceived differences in visual channels come from controlled models of human vision, which are generally constructed using large and visually isolated stimuli under optimal conditions. Visualizations, in contrast, often consist of large numbers of small marks viewed using a wide range of devices and environments. The assumptions made in controlled models of human vision may limit the utility of applying perceptual models to visualization design in practice.

These limitations are especially detrimental for color encodings. Environmental factors, display settings, and properties of visualization design can all inhibit people's abilities to distinguish encoded colors in visualizations [41, 50, 51]. Conventional color difference metrics, such as CIELAB, do not account for these factors, instead assuming large uniform color patches viewed in isolation under perfect conditions (2° or 10° of visual angle, approximately 50 pixels and 250 pixels wide for a standard web observer). However, many visualization systems rely on CIELAB and similar metrics to construct encodings, leading them to systematically underestimate the perceived differences between colors [41, 51]. This underestimation can lead to ineffective encoding choices by, for example, mapping continuous data to too narrow a range or encoding ranked or categorical data with colors that are too close together. Our goal is to provide a preliminary understanding of how we might adapt existing color difference models to account for visualization design factors.

We present a series of crowdsourced experiments used to model color difference perceptions for visualizations parameterized according to the designer's desired level of discriminability and known properties of a visualization. These models provide the first steps towards visualization-specific models of color difference, focusing on three different mark types: points, bars, and lines. Our models are grounded in comparisons of color mark pairs in a field of grey distractor marks in conventional visualizations (scatterplots, bar charts, and line graphs). Our experiments leverage an empirically-validated method from color science for constructing probabilistic models of difference perceptions to generate data-driven metrics for designers to consider when creating, evaluating, and refining visualizations. Our results show that conventional color difference metrics significantly underestimate the necessary differences between encoded values and that necessary differences between marks vary with the kind of visualization being used. For example, color encodings on elongated marks, such as those used in bar charts and line graphs, are significantly more discriminable than equally thick point marks, such as scatterplot points. The resulting models can be used to design color encodings with probabilistic bounds on their

• *Danielle Albers Szafir is with the University of Colorado Boulder. E-mail: danielle.szafir@colorado.edu.*

Manuscript received 31 Mar. 2017; accepted 1 Aug. 2017.
Date of publication 28 Aug. 2017; date of current version 1 Oct. 2017.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2017.2744359

1077-2626 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

predicted effectiveness based on known parameters of a visualization and to evaluate and refine existing codings.

Contributions: The primary contributions of this paper are empirical measures and models that capture how color differences vary across different mark types commonly used in visualizations (points, bars, and lines). These measures are made actionable through small modifications to CIELAB ΔE, a common color distance metric used in visualizations. Our results provide a quantitative analysis of color difference perceptions grounded in common visualizations, finding that perceived color difference varies inversely with size and that colors are more discriminable on elongated marks (bars and lines) than on points. They also formalize the relationship between elongation and perceived color difference. Integrating contextual information that influences encoding perceptions into visualization design tools through our models may significantly increase visualization effectiveness.

2 Background

Color provides three channels commonly used to visualize data: lightness, saturation/chroma, and hue [5]. While a number of tools and guidelines exist to guide effective color encodings (see Zhou & Hansen [55], Kovesi [23], and Silva et al. [47] for surveys), these metrics rely predominantly on heuristics, designer intuitions, and results from color science. As mentioned above, color science models are intended to capture parameters of human color vision [16], and do so by modeling vision agnostic of many of the complexities that may inhibit people's abilities to discriminate between colors in practice. Our work draws on previous studies of color in visualization and findings from color science to form the basis for our analyses.

2.1 Color in Visualization

Several studies in visualization have explored different aspects of color use. For example, studies have evaluated the impact of individual factors in colormap design, such as binning [40], color naming and semantics [20, 27], categorical similarity [15], cognitive bias [48], and salience [26]. Other studies measure the effectiveness of color encodings for different kinds of tasks. Cleveland & McGill [13] show that different color channels communicate individual values less precisely than position or size; MacEachren et al. [30] compare the effectiveness of color and other channels for investigating uncertainty. More recent work in graphical perception has identified potential utility of color for statistical judgments across large collections of datapoints [2, 3, 14]. In all of these studies, leveraging color effectively requires selecting color maps that span an appropriate encoding range to highlight important differences in the distribution.

In addition to formal studies on color use and application, a number of systems exist for creating encodings based on perceptual and cognitive heuristics (see Zhou & Hanseon [55] for a survey). Several of these systems use anticipated analysis tasks to guide color encoding design. For example, PRAVDAColor [4] introduces design heuristics that guide color encoding based on perception, task, and data type. Tominski et al. [52] design color encodings that explicitly enable comparison tasks. ColorCAT [34] expanded on this work to also consider localization and identification tasks as well as colorblindness.

Color map tools have also leveraged algorithmic processes over perceptual color spaces to generate encodings, such as harmonies [1], clustering [21], and statistical sampling [19]. ColorBrewer [18], among the most common color encoding tools currently in use, arose from a series of studies on the Hunt Model for mitigating color contrast in choropleth maps [8, 9]. The ramps emerging from these studies were then hand-tuned for aesthetics and performance. Subsequent studies provide algorithmic approximations of the resulting encodings by interpolating in CIELUV [54]. Colorgorical [17] optimizes across perceptual distance, nameability, and aesthetic considerations to generate categorical palettes.

While these tools provide visualization designers with access to encodings that abstractly support heuristic and perceptual constraints for effective color encodings, they are designed to work well in the abstract, based on color spaces modeled over small numbers of large, uniform marks viewed in isolation. Many visualization systems leverage small, non-uniform marks, such as scatterplot points or lines in a line graph, that may reduce the effectiveness of these encodings in practice (c.f., Fig. 1). We build on existing color spaces commonly used in visualizations to generate data-driven metrics for understanding how the effectiveness of these designs might change for different marks and visualizations.

2.2 Color Difference Metrics

Color difference metrics normalize color space such that the geometric and perceptual differences between colors are aligned (see Robertson [43] for a survey). For example, in CIELAB, one unit in color space approximately corresponds to one just-noticeable difference (JND). While these metrics vary significantly in how they define the set of available colors, they all characterize the psychophysical capabilities of human color perception. As a result, they rely on a simplified model of the world that allows these models to isolate the capabilities of the human visual system from the complexities introduced by real-world viewing [16].

This "simple world" assumption, while necessary for understanding the visual system, means that these models do not account for the complexities of real world viewing. For example, properties of the viewing environment, such as increased direct or ambient lighting [7, 39, 42], background and surrounding colors [38, 49], the size of a mark [12, 50], and display device [24, 45] may alter color perception in practice. This complicates the use of color difference metrics in visualization as these metrics are likely to underestimate the necessary differences between mark colors.

Our studies use CIELAB [25], a color space comprised of three primary axes: L^* (lightness), a^* (the amount of red or green), and b^* (the amount of blue or yellow). While metrics such as CIE ΔE_{94} [32], CIEDE2000 [29], and CIECAM02 [36] provide more precise calculations of color difference, the computational simplicity of CIELAB ΔE, where perceived difference is measured using Euclidean distance, makes it a popular choice for visualization tools (e.g. [11, 22, 28, 53]). Some encodings account for imperfections in CIELAB by hand (e.g. Samsel's green-blue encodings [44]), but this process does not scale well nor does it provide quantitative guarantees of effectiveness. Instead, this work focuses on providing actionable guidance for reasoning about perceptual effectiveness by generating quantitative metrics from existing models of color perception to aid in creating and evaluating encodings for visualizations.

2.3 Color Difference in Practice

Recent efforts provide methods for creating more robust models of color perception using crowdsourcing. For example, Szafir et al. [51] use a binary forced choice comparison between color patches to construct probabilistic models of color differences based on CIELAB. Stone et al. [50] extend that model to consider uniform marks of different sizes; however, these models consider only isolated colored squares, devoid of the visual complexity and shape variety found in visualizations. Reinecke et al. [41] ask participants to analyze gapped circles to compute discriminable differences and provide a thorough analysis with respect to different viewing factors. These three models show that conventional color difference models underestimate color perceptions in practice by roughly a factor of five. However, all of these models still fail to consider many of the complexities involved in visualizations, such as visual complexity, potential contrast effects [35], and varying mark shapes [12]. This work instead aims to create difference models specifically tailored to visualizations. We focus on properties known at design time, such as minimum bar or point widths. The results allow us to start to think about how design tools might account for complex perceptual phenomena in advance for visualization.

3 Visualization Factors in Color Perception

Visualizations violate the assumptions of conventional color science models in three ways:

1. **The Simple World Assumption:** Color science models assume perfect viewing conditions, whereas visualizations can be viewed anywhere (see Szafir et al. [51] for a discussion).

2. **The Isolation Assumption:** Color science models assume viewers are comparing one or two isolated color patches, whereas visualizations map colors in a visually complex environment.

3. **The Geometric Assumption:** Color science models assume all color patches are the same size and shape (generally either 2° (approximately 50 pixels wide) or 10° (approximately 250 pixels) wide), whereas visualizations map colors to marks of varying size and shape.

While prior models have addressed the Simple World Assumption [41,50,51], the other two assumptions are far less understood. The goal of this paper is to provide preliminary steps towards color difference models that account for all three violations simultaneously. In doing so, we can provide visualization designers with concrete probabilistic guidance that enables them to make more informed decisions about their color encodings grounded in what viewers will likely perceive.

In this work, we address the Simple World Assumption through crowdsourcing, using the methodology established in Szafir et al. [51]—a binary forced-choice comparison of two colored patches. We use this methodology as it has a natural analogue to comparing two datapoints in a visualization (asking people to identify whether two values are the same or different) and has produced comparable results to more complex methods (e.g., gapped circles [41]). We address the Isolation Assumption by mapping our test colors to marks in visualizations with a series of mid-gray distractor marks to add visual complexity more reflective of traditional visualizations (see §4.1 for details). Our studies do not offer the full visual complexity of traditional visualizations, which generally have large numbers of differently colored marks. However, the choice to use mid-gray distractor marks mitigates potential confounds from color contrast and also simplifies the comparison task by allowing participants to focus on the "colorful marks."

The Geometric Assumption, however, is more complicated to address. "Size" can have many meanings in visualization. Stone et al. [50] demonstrated that the diameter of an object can significantly change how readily we can discriminate between their colors, but tested only uniform squares. To date, the factors of size that directly effect color perceptions have yet to be enumerated. For example, using either length or area to encode data directly influence the size of a mark. Preliminary evidence shows that how we manipulate the size of a mark may have different ramifications on color perceptions [12], but these measures come from studies of response time in a visual search task. Instead, we consider the size factors in the context of marks applied to a visualization and the resulting effects of mark size on viewers' abilities to compare these marks.

To capture the possible variations from mark size, we first enumerated the different ways that mark size varies with values in a visualization (e.g., thickness, diameter, length, arc length, linearity, and area). Based on these enumerations, we identified four primary categories of marks that might affect color perceptions.

- **Diagonally Symmetric Marks:** Marks that have an equal height and width (e.g., points in a scatterplot, cells in a heatmap, bubbles in a cartogram).

- **Elongated Marks:** Marks that encode data using length in one dimension, but are fixed along the all others (e.g., bars in a bar chart, arcs in a donut chart).

- **Asymmetric Marks:** Marks whose length changes based on the position of its internal points but have a fixed thickness (e.g., lines in a line graph, arcs in a connected scatterplot, lines in a parallel coordinates plot).

- **Area Marks:** Marks that communicate information using their total area rather than any specific dimension of a mark (e.g., areas in a streamgraph, wedges in a pie, regions in a choropleth map).

These categories provide a basic scaffold for considering how we might build and reason about color and size in visualization. Specifically, we anticipate that color perceptions will vary individually across each class of mark size. To test these, we chose a canonical chart type for each category (scatterplots for diagonally symmetric marks, bar charts for elongated marks, and line graphs for asymmetric marks) except area marks to serve as our experimental stimuli. Area marks tend to vary quite irregularly (e.g., there is no fixed dimension). As a result, it is difficult to generate area marks that are both ecologically valid and cover the range of potential geometries. Additionally, since these marks often have a minimum area of 0 pixels, it is difficult to put probabilistic bounds on their visibility. In this study, we consider the relationship between different size factors associated with area marks (e.g., length and thickness). We anticipate our results will provide preliminary metrics that can be used for area marks, especially for visualizations such as choropleth maps and heatmaps, where the relative sizes of marks can be approximated in advance; however, a formal model for area marks is important future work.

4 General Methods

We compared color difference perceptions for points, bars, and lines using a series of mixed-factor experiments conducted on Amazon's Mechanical Turk. Each experiment focused on one mark type (points, lines, or bars) encoded as part of a visualization (scatterplot, line graph, or bar chart). All three experiments shared the same general structure, with variations to these experiments discussed in their respective Methods sections. Across all three experiments, the primary dependent measure was discriminability rate (how often color differences were perceived) and independent variables were mark size, color difference, and tested axis (L^*, a^*, and b^*).

4.1 Stimuli

We measured color difference perception using static visualizations rendered using D3 [6] (c.f., Fig. 4.1). Bar charts and scatterplots were rendered on a 375×250 pixel white background using 1 pixel mid-gray axes. Lines were rendered as 300×300 pixel visualizations to accommodate vertical spacing. Each stimulus contained two colored test marks and a series of randomly-placed distractor marks. Test marks were separated by 5° of visual angle (125 pixels) edge-to-edge. As the spatial distance between marks can influence color difference perceptions [10], we elected to preserve a constant distance between test marks. While visualization marks are generally variably spaced, using a constant mark distance mitigates potential confounds in our models due to mark spacing to provide an amortized prediction of color difference. We selected 5° (125 pixels) of separation as it provided a comfortable distance in piloting, approximately corresponds to the edges of foveal vision [33], and fits within most visualizations. Distractor marks were included to increase the visual complexity of the stimuli for increased ecological validity in our color comparison task and to address the Isolation Assumption. Test and distractor marks were mapped to a constant size, with size ranges for each experiment discussed in their respective Methods sections.

We sampled mark sizes along uniform steps in visual angle, allowing us to compare our results with those from color science. However, converting from visual angle to pixels requires knowledge of both how far the viewer is from the display and the display resolution, neither of which are generally available to visualization designers. As a result, all size conversions assumed a standard viewing distance of 30 inches, a D65 whitepoint,[1] and the HTML default pixel resolution of 96 dpi. [2] In most browsers, this will be remapped automatically to compensate for the actual display resolution. Tested sizes were first selected based on their appropriateness for the tested visualization type and then refined in piloting.

We mapped test marks to two different colors: one target color and a second color adjusted from the target color by a fixed color

[1] Converted using D3's CIELAB conversion modified to remove rounding
[2] http://www.w3.org/TR/css3-values/#absolute-lengths

Fig. 2. We tested color difference perceptions at six fixed color differences sampled according to the model constructed by Stone et al. [50]. This figure shows the $0.5ND_{(50,Size)}$, $1ND_{(50,Size)}$ (i.e., 1 JND), $1.5ND_{(50,Size)}$, and $2ND_{(50,Size)}$ (i.e., 2 JND) levels for L^* and 2° marks.

difference step. Target colors were computed by first uniformly sampling the CIELAB gamut from $L^* = 20$ to $L^* = 80$ in $12.5\Delta L^*$ and $12\Delta a^*$ and Δb^* steps. We then discarded all grays to avoid confusion with the distractor marks as well as all colors that would fall outside of the CIELAB gamut when adjusted by the largest tested amount. This resulted in 79 test colors ranging from $L^* = [30,65]$, $a^* = [-36,48]$, and $b^* = [-48,48]$. We computed adjusted colors by sampling $\pm$ 6 steps along each axis from the tested color. Step sizes were computed by interpolating the $ND_{(50,size)}$ model from Stone et al. [50], a size-sensitive crowdsourced model where $1ND_{(50,size)}$ corresponds to the color difference in CIELAB we expect will be detected 50% of the time. Stimuli used steps of $0.5ND_{(50,Size)}$, $0.75ND_{(50,Size)}$, $1ND_{(50,Size)}$, $1.25ND_{(50,Size)}$, $1.5ND_{(50,Size)}$, or $2ND_{(50,Size)}$ (c.f., Fig. 2). We mapped all distractor marks mid-gray ($L^* = 50$) to minimize any potential conflicts from simultaneous contrast.

4.2 Procedure

All three experiments used a binary-forced choice design, asking participants to report whether two colored marks appeared to be the "same" color or "different" colors. Participants completed the study within their web browser with each stimulus rendered in real-time.

Participants were first screened for color vision deficiencies using four Ishihara plates. Because consumer monitors are likely uncalibrated, the Ishihara plates provide an approximate screening. We supplemented this screening by asking participants to self-report any color vision deficiencies. They were then shown three example stimuli to clarify the definition of "same" and "different" colors: one stimuli had identically colored marks, a second had marks of differing hues, and a third had marks of differing lightnesses. Participants had to correctly complete all three tutorial questions before beginning the study.

Participants then completed a series of 79 comparisons (one for each test color), reporting whether test marks appeared to be the same color or different colors using keyboard inputs ('f' for same, and 'j' for different, consistent with Stone et al. [50]). The trial window persistently showed input keys above the stimuli on the corresponding sides of the display ('f' on the left, 'j' on the right). Each trial randomly mapped the 79 test colors to different conditions, and conditions were presented in a random order to mitigate learning and fatigue effects. To mitigate contrast effects between subsequent trials, participants saw a $0.5s$ gray adaptation screen between each trial. We concluded by collecting demographic information. We included four same-color stimuli to mitigate bias and three large-difference stimuli (nameably different colors greater than $20\Delta E$ at middle sizes) an engagement check to ensure honest participation. Participants incorrectly responding to two or more large-difference stimuli or averaging less than $0.5s$ (grey screen duration) per response time were excluded from our analysis. Across all experiments, color difference was a within-participants factor, while tested axis (L^*, a^*, or b^*) was a between-participants factor.

4.3 Analysis & Modeling

The primary dependent measure in these experiments is the *discriminability rate*—the proportion of trials where a difference is correctly recognized (i.e., $\frac{reported\ differences}{total\ trials}$)—for each combination of independent variables. We first analyzed discriminability rates from each experiment using an ANCOVA to identify factors influencing difference perceptions. Because prior studies found small effects of sample color [50] and crowdsourcing may introduce unexpected individual differences, sample color and between-participant variation were treated as random covariates. These covariates help mitigate incidental effects from crowdsourcing and small biases introduced by the tested color sample [37]. We perform post-hoc comparisons using Tukey's Honest Significant Difference Test (HSD).

We then constructed a model of color difference perceptions parameterized by mark size using the approaches described in Szafir et al. [51] and Stone et al. [50], which renormalize CIELAB ΔE based on discriminability rates collected across each axis to account for axis-level variation. This model provides a validated metric for generating controlled models of perceived color difference using a relatively small number of samples and is designed for use on Mechanical Turk. The model computes $\Delta E_{p,s}$ as follows:

1. Preprocessing: We first compute the discriminability rate for each combination of size and color difference, with our sampling rate predicting a 7% margin of error based on a 50% discriminability rate. We then verify all rates fall below the asymptote (or "knee" [12]) in the resulting probability curve, which represents the threshold at which differences are immediately perceptible. All data collected in our studies fell within this range.

2. Size $\times$ Axis Models: We model the resulting discriminability rates for a given mark size and axis in CIELAB using linear regression, treating discriminability rate as our dependent variable, distance (ΔE) between colors as our independent variable, and between-participant variation and starting color as random effects. We constrain the regression to a zero-intercept to account for small variations due to our sampling methodology. The resulting model has the form:

$$p = m_x * \Delta x \tag{1}$$

where p is the proportion of detected color differences ($p = 50\%$ is a typical JND), m is the regression line slope, and x is the current CIELAB axis. We can alternatively compute the color difference necessary to achieve a $p\%$ noticeable difference ($ND(p)$) in ΔE as:

$$ND_x(p) = \frac{p}{m_x} \tag{2}$$

3. Size-Independent Models: We model size variation as a function of the set of slopes m_x as an inverse function of size s. Combined with Eqn 2, we model a $p\%$ noticeable difference as:

$$ND_x(p,s) = \frac{p}{c_x + \frac{k_x}{s}} \tag{3}$$

where c and k are constants derived from a linear fit of slopes to inverse size. The resulting model corresponds to a quantitative bound on the minimal discriminable color difference in CIELAB ΔE for each mark type based on the geometric properties of that mark. We report our models as both a function of p and as 50% JNDs ($ND(50\%,s)$).

4. Normalized Color Difference ($\Delta E_{p,s}$): We can compute $ND_x(p,s)$ for each axis of CIELAB and divide each term in the resulting Euclidean distance model to normalize CIELAB according to the anticipated visualization design. We express this normalized $\Delta E_{p,s}$ as:

$$\Delta E_{p,s} = \sqrt{(\frac{\Delta L}{ND_L(p,s)})^2 + (\frac{\Delta a}{ND_a(p,s)})^2 + (\frac{\Delta b}{ND_b(p,s)})^2} \tag{4}$$

which predicts that a color difference of $\Delta E_{p,s} = 1.0$ will be detectable by $p\%$ of viewers for a given mark size. We use the resulting model to compute size-scaled differences across the full color space, including inter-axis variation.

While this modeling methodology appears fairly simple, it has been directly validated in the color science community where it predicted cross-axis 50% JNDs to within 1% for crowdsourced users [51], has produced consistent in subsequent studies [50], and generated comparable JNDs to alternative methodologies [41]. Data tables used in our calculations and our experimental infrastructure can be found at `http://cmci.colorado.edu/visualab/VisColors`.

Table 1. Regression results for points, where $p = m_x * \Delta X$.

Axis	Size (s)	Size in Px	Slope	R^2	$ND(50\%)$ in ΔE
L	0.25°	6 px	0.059	0.948	8.37
L	0.5°	12 px	0.074	0.97	6.74
L	0.75°	18px	0.087	0.981	5.75
L	1°	25px	0.087	0.965	5.75
L	1.5°	37 px	0.082	0.996	6.08
L	2°	50 px	0.091	0.974	5.47
a	0.25°	6 px	0.031	0.984	16.11
a	0.5°	12 px	0.05	0.988	9.98
a	0.75°	18px	0.059	0.987	8.52
a	1°	25px	0.064	0.992	7.81
a	1.5°	37 px	0.073	0.985	6.87
a	2°	50 px	0.073	0.994	6.84
b	0.25°	6 px	0.026	0.978	19.46
b	0.5°	12 px	0.037	0.988	13.34
b	0.75°	18px	0.044	0.994	11.35
b	1°	25px	0.05	0.979	10.03
b	1.5°	37 px	0.056	0.979	8.97
b	2°	50 px	0.063	0.99	7.99

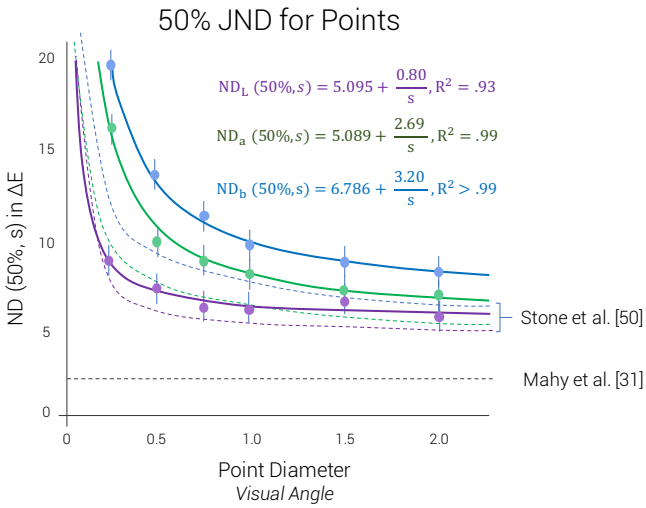

Fig. 3. We can use our data to model 50% JNDs ($ND(50\%,s)$) for scatterplots. JNDs shrink significantly as point marks grow larger. Even near the asymptote, JNDs for scatterplot points on the web are significantly larger than those predicted in laboratory studies (e.g., Mahy et al. [31]) and non-visualization studies (e.g., Stone et al. [50]). Error bars represent expected margin of error from our sample size.

4.4 Participant Recruitment

One of the primary goals of this work is to understand color perceptions in the context of visualization viewing. As a result, we want to balance ecological validity with controlled quantitative modeling. While viewing environment can significantly impact color perceptions, visualizations are generally viewed in imperfect environments in practice. However, recent work [41,51] has shown that sampling across this variation can generate accurate models of color perceptions in practice. To construct color difference models that are robust to anticipated viewing variations, we recruited participants for our studies using Amazon's Mechanical Turk.

Across all three studies, we recruited 461 total participants from the U.S. Mechanical Turk population. All participants has a 95% or greater approval rating. 4 participants were excluded from our analysis due to self-reported color vision deficiencies, 18 due to poor performance on the large-difference stimuli (incorrectly identifying more than 2 of the 4 nameably different colors), and 7 for mean response times less than 0.5 seconds.

5 Experiment One: Scatterplots

Scatterplots generally use small, diagonally symmetric marks to encode data. These marks most closely align with the uniform fields used to construct conventional color science models. As a result, we hypothesized that existing models of color perceptions, especially the color-size model presented in Stone et al. [50], would provide a reliable model of color difference perceptions for scatterplots.

5.1 Methods

To model color difference for diagonally symmetric points, we generated a series of scatterplots with circular marks. Mark diameters ranged from 0.25° (6 pixels) to 2.0° (50 pixels). Test points were mapped to a random y-value and separated by 5° (125 pixels) along the x-axis. We computed the distractor positions by placing $\sqrt{\frac{visualization\,area}{mark\,area}}$ points at positions randomly sampled from a normal distribution ($\mu = 0.5, \sigma = 1.0$) and removing points intersecting any other point in the plot. Table 5 summarizes the tested sizes, and Figure 1 shows examples of the smallest and largest stimuli.

We ran a 6 (diameters, within) × 6 (color differences, within) × 3 (color axis, between) mixed-factors experiment to collect data for our model. Each participant saw each diameter × color difference combination twice plus seven engagement checks. Each stimulus used a random test color and was presented in a random order.

5.2 Results

We collected data from 81 participants. Data from seven participants was excluded for poor performance on large-difference stimuli, and two participants were excluded for response time, resulting in 72 participants (30 female, 41 male, 1 did not report, $\mu_{age} = 34.3 \pm 10.5$; 5,688 total samples, 24 participants per axis).

5.2.1 Analysis of Factors

We analyzed the frequency of reported differences across experimental (axis, diameter, and question order) and demographics (gender, display type, and age) factors using a six-level ANCOVA with starting color treated as a random covariate. We found a significant effect of axis on discriminability ($F(2,61) = 4.7, p < .02$), with participants identifying differences significantly more for L^* stimuli ($\mu_L = 74.9\%$) than a^* or b^*($\mu_a = 61.7\%; \mu_b = 60.2\%$), consistent with prior work [43,51]. We found no transfer (e.g., learning or fatigue), gender, or display effects. We did find a marginal effect of age ($F(1,43) = 5.9, p < .06$), but no systematic bias ($R^2 = .003$).

5.2.2 Modeling

We first computed the discriminability rates for each of the 72 combinations of axis × mark diameter × ΔE. We modeled the collected data using the procedure discussed in §4.3. We first fit a linear regression to the discriminability rates and forced the regression through 0 to mitigate small fluctuations introduced by our data-driven approach. All models fit with $R^2 > 0.94$ (Table 5). We then modeled these slopes as an inverse function of mark diameter. The resulting models were:

$$ND_L(p,s) = \frac{p}{0.0937 - \frac{0.0085}{diameter}} \quad (5)$$

$$ND_a(p,s) = \frac{p}{0.0775 - \frac{0.0121}{diameter}} \quad (6)$$

$$ND_b(p,s) = \frac{p}{0.0611 - \frac{0.0096}{diameter}} \quad (7)$$

with $L^*(F(1,4) = 35.56, p < .004, R^2 = .90)$, a^* $(F(1,4) = 132.47, p < .0003, R^2 = .97)$, and $b^*(F(1,4) = 35.01, p < .005, R^2 = .90)$ models all providing statistically significant fits to the data. Designers can select their desired minimum sizes and JND levels (p) to compute $ND_{p,s}$ for each axis. These $ND_{p,s}$ values plug into Eqn. 4 to renormalize CIELAB for scatterplot points, considering distance along any combination of axes. Figure 3 plots the 50% JNDs ($ND(50\%,s)$)

models generated from our data against those from the individual regressions computed for each combination of size and axis.

Our results confirms findings from prior studies: marks become less discriminable as their size decreases and real devices have significantly larger JND thresholds than those from traditional color science metrics. We replicated previous results that show **perceptible color differences for points vary inversely with point diameter**. However, we also find that our results predict larger JNDs for scatterplots than the isolated marks modeled in Stone et al. [50] (Fig. 3). For example, a 0.5° mark in our scatterplot stimuli has an a^* axis JND of $10.47\Delta E$, whereas isolated mark JNDs would require $8.42\Delta E$. While some of this variation may be explained by margin of error due to sampling, we see this bias systematically across all three axes. We anticipate that the increased JNDs are likely due to the increased visual complexity associated with visualizations—viewers have additional visual information that may complicate data interpretation—suggesting effects due to the Isolation Assumption. Future testing is needed to understand how this complexity might influence these judgments.

6 Experiment Two: Bar Charts

Elongating marks decreases the time taken to identify marks of different colors [12], suggesting that elongated marks are easier to differentiate. As a result, we hypothesized that our point-based model would *overestimate* perceived color differences for bars and other elongated marks, causing designers to be overly conservative in their color choices and superficially reducing the range of available encoding colors. We can model difference perceptions for elongated marks to provide benchmarks for evaluating color encodings for elongated marks in presentation-based visualizations where data is known at design time. We also explored whether designing for known mark thickness (the length of the fixed edge, horizontal width in our study) may provide sufficient discriminability thresholds for designers when data (and, consequently, bar length) is unknown.

6.1 Methods

Measuring color difference perception for bars required considering not only mark thickness, but also mark length. We sampled bar thickness from 0.25° (6 pixels) to 2° (50 pixels), and bar lengths from 0.125° (3 pixels) to 6° (150 pixels). In piloting, the increased number of conditions led to significant fatigue effects. To mitigate these effects, we treated thickness and length as mixed-participants factors, with each participant seeing three thicknesses and four lengths.

We used vertical bar charts with test marks separated on the x-axis by 5° (125 pixels) and distractors uniformly spaced between the test marks as well as between each test mark and the bounds of the visualization. Distractor length were randomized in the range 0.125° (3px) to 6.0° (150px). Fig. 1 shows examples of the resulting plots.

We ran a 6 (thicknesses, blocked between) × 8 (lengths, blocked between) × 6 (color differences, within) × 3 (color axis, between) mixed-factors experiment to collect data for our model. Lengths and thicknesses were blocked between participants, with each participant seeing thicknesses of either 0.25°, 0.75°, and 1.5° or 0.5°, 1.0°, and 2.0° and lengths of either 0.125°, 0.5°, 1.0°, and 3.0°, or 0.25°, 0.75°, 1.5°, and 6.0°. Each participant saw each combination of three thicknesses and four lengths once for each color difference plus the seven engagement check stimuli, with each stimulus mapped to a random test color and presented in a random order.

6.2 Results

We recruited 301 participants for this study. We excluded data from four participants due to self-reported CVD, six for poor performance on the large-difference stimuli and three for response time, resulting in 288 participants (141 female, 146 male, 1 DNR, $\mu_{age} = 34.9 \pm 10.2$; 22,752 samples, 24 participants per axis × thickness × length).

6.2.1 Analysis of Factors

We analyzed reported differences across experimental (axis, thickness, length, and question order) and demographics (gender, display type, and age) factors using a seven-level ANCOVA with starting color treated as a random covariate. We found a significant effect of axis on perceived difference ($F(2,280) = 5.14, p < .007$), with participants identifying differences significantly more for L^* axis stimuli ($\mu_L = 71.8\%$) than b^* ($\mu_b = 64.6\%$). We also found significant effects of both thickness ($F(1,280) = 112.65, p < .0001$) and length ($F(1,280) = 523.19, p < .0001$), suggesting that the Stone et al. [50] model used as a baseline did not predict variation with size for bars. We found no evidence of transfer (e.g., learning or fatigue), gender, or display effects. We did find a significant effect of participant age ($F(1,280) = 6.28, p < .02$); however, the effect did not reveal any systematic bias ($R^2 = .003$).

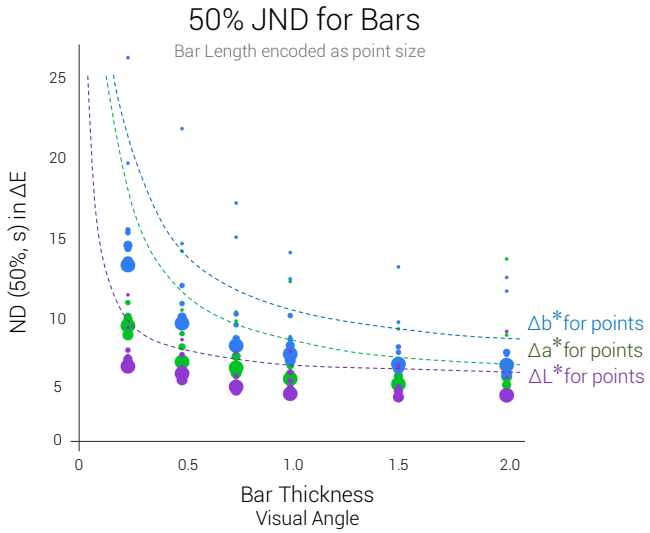

Fig. 4. Plotting thickness (x-axis), length (point size), and $ND(50\%, s)$ (y-axis), we see that longer bars correspond to lower JND thresholds for the same thickness. This behavior is asymptotic: as ratios between the length and thickness become larger (e.g., large points on the left), ND50 values cluster more tightly together.

6.2.2 Modeling

We first computed the discriminability rates for each combination of axis × mark thickness × mark length × ΔE. We modeled all collected data using the procedure discussed in §4.3. We first fit a linear regression to the discriminability rates and forced the regression through 0 to mitigate small fluctuations introduced by our data-driven approach. All models fit with $R^2 > 0.89$ ($\mu_{R^2} = .957, \sigma_{R^2} = .03$, data available on the project webpage). We found a significant fit along all three axes between model slope and bar thickness and length:

$$ND_L(p,s) = \frac{p}{0.1061 - \frac{0.0107}{thickness} - \frac{0.003}{r}} \quad (8)$$

$$ND_a(p,s) = \frac{p}{0.0895 - \frac{0.0111}{thickness} - \frac{0.0037}{r}} \quad (9)$$

$$ND_b(p,s) = \frac{p}{0.0751 - \frac{0.0113}{thickness} - \frac{0.003}{r}} \quad (10)$$

where $r = \frac{length}{thickness}$ with $L^*(F(2,45) = 61.3, p < .0001, R^2 = .73)$, a^* ($F(2,45) = 76.2, p < .0001, R^2 = .77$), and b^* ($F(2,45) = 96.6, p < .0001, R^2 = .81$) models all providing statistically significant fits to the data. Figure 4 models $ND(50\%, s)$ computed from this data compared to the point models.

Colors were generally more discriminable on bars than on scatterplots of equal thickness. **Colors on longer marks were also more discriminable than on shorter bars of equal thickness**. Gains from this elongation were asymptotic, generally leveling off at gains of $5\Delta E$ to $10\Delta E$ (Fig. 4), with functions reaching an asymptote around a 2:1 ratio of length to thickness. Because our goals are to inform encoding

design, we specify our models as a function of thickness; however, mark size can be expressed through four parameters: area, elongation ($\frac{longest\ edge}{shortest\ edge}$), longest edge, and shortest edge.

We conducted a four-factor ANOVA for the model slopes to disentangle contributions of these parameters, testing both main effects and specific interactions between area and edge length and between elongation and edge length. We found main effects across all three axes for elongation ($F_L(8,39) = 9.93, p < .004$, $F_a(8,39) = 3.05, p < .09$, $F_b(8,39) = 14.62, p < .0005$) and longest edge ($F_L(8,39) = 5.85, p < .03$, $F_a(8,39) = 6.15, p < .02$, $F_b(8,39) = 15.6, p < .0003$), but no significant effect of area. We found interaction effects of elongation $\times$ longest edge ($F_L(8,39) = 3.28, p < .08$, $F_a(8,39) = 5.39, p < .03$, $F_b(8,39) = 3.8, p < .06$) and elongation $\times$ shortest edge ($F_L(8,39) = 7.41, p < .001$, $F_a(8,39) = 3.11, p < .09$, $F_b(8,39) = 6.42, p < .02$).

From these findings, a more precise model for elongated marks can be expressed as a function of the shortest edge using:

$$ND_L(p,s) = \frac{p}{0.1056 - \frac{0.0061}{shortest} - \frac{0.0134}{elongation}} \tag{11}$$

$$ND_a(p,s) = \frac{p}{0.0881 - \frac{0.0067}{shortest} - \frac{0.0117}{elongation}} \tag{12}$$

$$ND_b(p,s) = \frac{p}{0.0719 - \frac{0.0059}{shortest} - \frac{0.0105}{elongation}} \tag{13}$$

where $elongation = \frac{longest\ edge}{shortest\ edge}$ with $L^*(F(2,45) = 52.9, p < .0001$, $R^2 = .70)$, a^* $(F(2,45) = 77.7, p < .0001$, $R^2 = .78)$, and b^* $(F(2,45) = 50.3, p < .0001$, $R^2 = .69)$ models all providing statistically significant fits to the data. It is important to note that these edge models are limited in their statistical power as we did not collect data at evenly distributed elongation ratios. Future work should collect additional data at specific elongation ratios to validate this model.

The magnitude of the discriminability gains of elongated marks suggest significant potential benefits for designers in considering mark shape as part of their design processes. Considering the limited space available for encodings (L^* ranges from 0 to 100), gains from elongation provide designers with significantly more encoding space than point-models alone. Our findings imply that even when bars are wider than they are tall, we can leverage our models to predict a conservative baseline color difference threshold using a mark's minimum thickness. However, predictive guidance based on fixed bar thickness would not be robust for very short vertical bars. In these cases, our models are best used for post-hoc validation. When we anticipate marks will be elongated, however, we can leverage these models for considerably more aggressive encoding practices to maximize the possible discriminable differences in data.

7 Experiment Three: Line Graphs

Experiment Two demonstrated that elongated bar marks increase perceived color differences compared to diagonally symmetric point marks. However, significantly elongated marks, such as lines, exceed the asymptotic discriminability behaviors seen in Experiment Two and are often visually asymmetric, curving according to parameters of the data. Therefore, we can measure perceived color difference on lines to generate models for color encoding design for marks beyond these asymptotic bounds, allowing designers to maximize their use of the encoding space based on prescribed line thickness. We anticipate that lines will be significantly easier to discriminate than diagonally symmetric points of equivalent thickness.

7.1 Methods

We tested color difference perception using six different line thicknesses, ranging from 0.05° (1 pixel) to 0.35° (9 pixels, Table 2). We tested a smaller size range than with scatterplots as line thicknesses tend to be much smaller than the diameter of scatterplot points. We generated test marks by plotting 38 y-values randomly sampled between 0 and 75 from a normal distribution. We plotted these values at uniform intervals along the x-axis (300px) and then interpolated the resulting points using a Bezier curve. To preserve a 5° spacing between the test marks, one test mark was drawn between $y = 0$ and $y = 75$, and the second was drawn between $y = 200$ and $y = 275$. To make the line comparison task more natural, we juxtaposed marks vertically, rather than horizontally. While we do not anticipate any confounds from this choice, better understanding the impact of orientation on comparison tasks is important future work. Figure 1 shows examples of the resulting plots for the largest and smallest line thicknesses. Distractor lines were constructed using the same procedure, but mapped to a random y position. Test marks were always rendered on top of the distractor marks to avoid occlusion.

The models we used to compute color difference steps in Experiments One and Two suggest very large step sizes for small marks (e.g., $ND_{b^*}(50\%, 0.05^\circ) = 63\Delta E$). However, in piloting, we found that much smaller color difference steps were often discriminable. As a result, all thicknesses below 0.25° (6 pixels) were mapped to the same color difference step sizes as 0.25° marks.

We ran a 6 (thicknesses, within) $\times$ 6 (color differences, within) $\times$ 3 (color axis, between) mixed-factors experiment to collect data for our model. As with scatterplots, each participant saw each combination of size and color difference twice, using a random test color, and presented in a random order.

7.2 Results

We recruited 79 participants for this study. We excluded data from 5 participants for poor performance on the large-difference stimuli and 2 for response time, resulting in 72 participants (29 female, 43 male, $\mu_{age} = 34.1 \pm 11.0$; 5,688 trials, 24 participants per axis).

7.2.1 Analysis of Factors

We analyzed the frequency of reported differences across experimental (axis, thickness, and question order) and demographic (gender, display type, and age) factors using a six-level ANCOVA with target color treated as a random covariate. We found a significant effect of axis on perceived difference ($F(2,61) = 5.2, p < .001$), with participants identifying differences significantly more frequently for L^* axis stimuli ($\mu_L = 62.0\%$) than a^* or b^* ($\mu_a = 77.9\%; \mu_b = 77.4\%$). We also found a significant effect of line thickness ($F(1,5) = 294.3, p < .0001$), suggesting that point-scale ΔE approximations did not adequately capture JND across all line thicknesses. We found no evidence of transfer, gender, age, or display effects.

7.2.2 Modeling

We first computed the discriminability rates for each of the 72 combinations of axis $\times$ line thickness $\times$ ΔE. Our regression models included all computed means and fit with $R^2 > 0.87$ (Table 2, $\mu_{R^2} = 0.94, \sigma_{R^2} = 0.02$). As with points, lines also fit to an inverse function of line thickness. The resulting models were:

$$ND_L(p,s) = \frac{p}{0.0742 - \frac{0.0023}{thickness}} \tag{14}$$

$$ND_a(p,s) = \frac{p}{0.0623 - \frac{0.0015}{thickness}} \tag{15}$$

$$ND_b(p,s) = \frac{p}{0.0425 - \frac{0.0009}{thickness}} \tag{16}$$

with L^* $(F(1,4) = 32.38, p < .005, R^2 = .89)$, a^* $(F(1,4) = 18.03, p < .02, R^2 = .82)$, and $b^*(F(1,4) = 13.94, p < .03, R^2 = .77)$ models all providing statistically significant fits to the data. Designers can select their desired minimum sizes and JND levels (p) to compute $ND_{p,s}$ for each axis to renormalize CIELAB using Eqn. 4 to compute differences across all three axes for lines.

Figure 5 plots the 50% JND ($ND(50\%, s)$) model for lines predicted by our model against those from the individual regressions and against the scatterplot models from Experiment One. We found that **perceptible color differences for lines vary inversely with thickness**, and lines are significantly more discriminable than equally thick points. However, the elongated areas provided by line graphs made marks

Table 2. Regression results for lines, where $p = m_x * \Delta X$.

Axis	Size (s)	Size in Pixels	Slope (m)	R^2	$ND(50\%)$ in ΔE
L	0.05°	2px	0.033	0.876	15.35
L	0.1°	3px	0.042	0.92	11.98
L	0.15°	4px	0.058	0.921	8.69
L	0.25°	6px	0.065	0.955	7.74
L	0.3°	7px	0.069	0.947	7.23
L	0.35°	9px	0.072	0.96	6.92
a	0.05°	2px	0.036	0.978	13.92
a	0.1°	3px	0.043	0.956	11.57
a	0.15°	4px	0.049	0.959	10.28
a	0.25°	6px	0.053	0.94	9.39
a	0.3°	7px	0.061	0.933	8.15
a	0.35°	9px	0.064	0.919	7.79
b	0.05°	2px	0.026	0.981	19.47
b	0.1°	3px	0.031	0.967	16.15
b	0.15°	4px	0.033	0.934	15.17
b	0.25°	6px	0.036	0.918	13.75
b	0.3°	7px	0.04	0.927	12.43
b	0.35°	9px	0.045	0.945	11.05

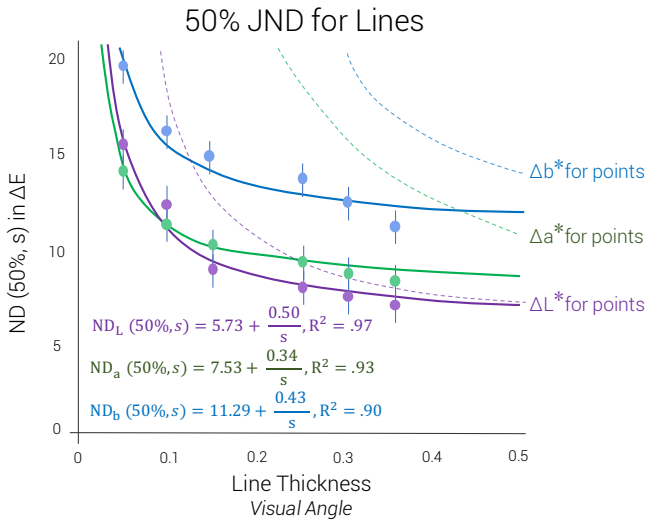

Fig. 5. Just noticeable differences between colors in a line graph shrink significantly as line thicknesses grow larger. Lines are generally more discriminable than equally wide scatterplot marks (dashed lines). Error bars represent expected margin of error.

significantly easier to discriminate than scatterplot points. Further, line thicknesses are less sensitive to variations in size, reflected in the smaller coefficients in Fig. 5. For short lines (those below the asymptotic edge ratio), our bar models provide a closer approximation of intended difference. However, most line marks far exceed this ratio. By designing according to perceptual thresholds for lines, designers can take advantage of the visual system's capabilities to better discriminate between elongated marks to help maximize use of available colors in visualizations. The asymptotic behavior of elongation allows us to provide more precise color difference metrics on asymmetric elongated marks.

8 Discussion

We provide preliminary steps towards a set of color difference models tailored to visualizations. Our findings address three challenges in leveraging models from color science for visualization:

- **The Simple World Assumption:** We replicate prior results that indicate color perceptions can be measured in crowdsourced environments to inform design practice [41, 50, 51].
- **The Isolation Assumption:** We measure color difference perceptions in simple visualizations that include distractor marks simulating possible data distributions.
- **The Geometric Assumption:** We compare color difference perceptions on point, bar, and line marks, systematically varying point diameter, bar thickness and length, and line thickness.

Our results suggest the importance of understanding color perceptions in the context of visualizations. First, our models confirm prior findings [50, 51] that traditional measures of color difference are not robust for real-world viewing: a 2° JND for a scatterplot point was roughly 3 times larger than that predicted in controlled environments [31]. By modeling perceived color difference as a function of the probability a difference be detected (p), we allow designers control over how robustly discriminable their encodings will be.

Second, our models allow designers to quantitatively reason about trade-offs between encoding range and the number of discriminable differences for visualizations. Effective color mappings have a limited set of perceptible colors to work with, but need to make important differences salient in the data. Our results indicate that just noticeable color difference generally varies inversely with the thickness of a mark and that elongated marks used in many visualizations (those where one edge is longer than the other, such as bars and lines) are more discriminable than to uniform marks with equivalent shortest edges. By leveraging known design properties of a visualization, such as minimum bar or line thickness, we can allow designers to quantitatively reason about the effectiveness of their encoding choices.

These models also allow us to identify limitations in existing encodings. For example, ColorBrewer is heavily used in visualizations, but its encodings were designed cartography, which generally use larger marks than scatterplots or line graphs [18]. As a preliminary proof of concept for applying our models to visualization design, we used our 50% JND models to evaluate ColorBrewer for scatterplot points and line graph lines approximating default Tableau parameters (10px diameter and 4px thickness respectively). We found that 13 of the 18 ColorBrewer 9-step sequential ramps were not robust to these mark sizes: only YlGn, YlGnBl, OrRd, YlOrBr, YlOrRd, and Reds retained at least 1 JND between subsequent steps. [3] This application suggests that visualization designers should closely consider size in selecting encodings: even designer-curated encodings may not be robust for many common visualizations. Some systems have begun to integrate similar constraints into their visualizations. For example, Tableau 10 [4] and d3-jnd [5] make use of size-dependent JND models. Size and mark models such as those presented here provide a preliminary basis for predicting robustness and refining encodings to help designers adapt to the perceptual constraints of visualizations. Our future studies will explore how robust popular encodings are for different visualization types and can verify our models' predictions against expert practices.

8.1 Limitations & Future Work

While our work significantly extends knowledge of color perception for visualization, several aspects of our experiments exchange ecological validity for modeling control. The limitations of this approach provide opportunities for future refinement of these models and new understandings of perception for visualization.

First, the visualizations themselves were constrained to support task simplicity. For example, only the target marks were colored, mitigating contrast effects present in real visualizations and simplifying the experimental task. The lack of contrast effects is also why we opted not to test heatmaps, which are generally heavily affected by contrast. Marks were tested at fixed distances to avoid potential confounds from variable distance comparisons [10]. Scatterplot points were aligned

[3] Details available at http://cmci.colorado.edu/visualab/VisColors

[4] https://www.tableau.com/about/blog/2016/7/colors-upgrade-tableau-10-56782

[5] https://github.com/connorgr/d3-jnd

to reduce time costs associated with visual search. We anticipate that our results have greater ecological validity than existing models, but some of these simplifications may lead to inaccurate predictions in some cases. Future work should extend these models to consider a more holistic set of design factors. Part of this work should validate the provided models by asking participants to compare datapoints in a broader variety of real-world visualizations with color, mark size, and data distributions drawn from real data and designer practices. This study would verify how robust the mark-specific models are to the simplifications made in these studies (e.g., the grey distractors and simple visualizations).

Second, the models were generated using a sampling of crowdsourced participants. While the limitations of this choice are discussed in detail in Szafir et al. [51], we anticipate that the use of crowdsourced data actually improves the validity of these metrics for design applications, especially as toolkits such as D3 [6] and Vega [46] increasingly simplify web-based visualization development. In prior studies, the methodology used in this paper improved predictive performance for the web from 14% accuracy with traditional models to 99% accuracy [51]. However, this method relies heavily on a representative sample of users and provides little insight into precise mechanisms of human vision due to noise introduced by variance in viewing conditions. Because our results align closely with those predicted in prior models (e.g., [41, 50, 51]), we believe our sample provides reliable metrics; however, future work could extend the results presented here to larger user samples, marks sizes, and specific types of devices. Further, as display technology changes, the models may need to be revised at regular intervals to accommodate new display parameters.

Finally, we elected to use CIELAB due to its common use in visualization, validated methodological use in past studies [51], and computational simplicity. Future work may consider the use of more complex color difference models, such as CIECAM02 [36]. We anticipate leveraging these spaces will result in more accurate and holistic considerations of color difference perceptions; however, the Euclidean renormalizations used here may not generalize well to more complex, piecewise metrics. Future work will need to consider how to construct data-driven adaptations of these models while still allowing efficient computation and application.

8.2 Using the Models for Visualization

Visualization designers can use our results to quantitatively reason about the effectiveness of their visualizations. We anticipate applications of our metrics for encoding design, assessment, and refinement. For example, designers often specify several properties of their intended visualizations *a priori*, such as mark types and the range of allowable mark sizes. These models allow designers to use known properties of a visualization to guide color encoding choices: designers can generate a threshold in CIELAB that identifies a minimum discriminability level computed from the smallest allowable sizes and elongation factors determined by the mark shape parameters. These thresholds enable *a priori* evaluation and refinement of predesigned ramps to predict their likely utility for the target design. These predictions allow designers to anticipate and account for potential limitations of candidate encodings in end designs.

Size parameters of certain mark types, such as areas and bar length, cannot be specified without access to the target data. In these cases, the specified models can be leveraged to instead verify when and how specific end visualizations might fail based on their color choices, similar to ColourCheck [41]. In such cases, our models provide significantly more precise estimates of perceived differences than existing systems, which generally do not consider small or non-uniform marks. For example, visualizations in data journalism generally use fixed datasets, and journalists can use our models to estimate how accurately their readers interpret color-encoded data.

In extensible visualization tools, where data is not fixed nor are its properties known *a priori*, designers can use these models to automatically refine their encodings by adapting the color ranges or bins to support a prescribed number of discriminable steps when the data is loaded. For example, when marks become smaller, systems might push the endpoints of an encoding further apart to preserve desired distances. Figure 6 shows the effects of increasing the distance between endpoint colors to new sizes using JND proportions from our scatterplot models. Because these models rely on simple modifications to Euclidean distance metrics, manipulating encodings based on desired difference thresholds can be done in real-time.

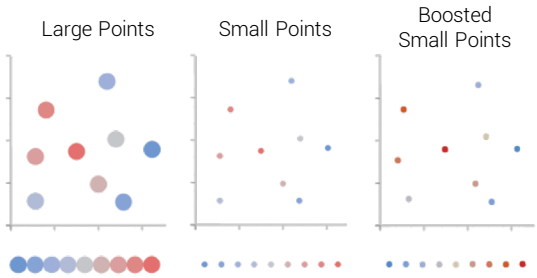

Fig. 6. We can use our models to automatically increase or decrease differences between colored marks, pushing colors from the large scatterplot apart proportionally to preserve relative differences for our small marks (right).

9 Conclusion

In this paper, we measure factors of and construct data-driven models for color difference perception in visualization. These models focus on three mark types—points, bars, and lines—modeled across different size parameters to generate a set of probabilistic metrics for guiding color encoding design and evaluation. We find that our models align well with previous in-practice models of color difference, but that elongated marks, as commonly used in visualizations, significantly increase discriminability over fixed-thickness models. We envision these metrics as first steps towards building a quantitative understanding of color perception in visualization and a broader discussion of the adaptability of controlled perceptual models to the complexities of visualization viewing in practice.

Acknowledgments

We thank Maureen Stone, Vidya Setlur, and Michael Gleicher, co-creators of the methodologies in [50, 51], for their prior support and mentorship and Daniel Szafir and the VIS reviewers for their feedback on this paper. This research was funded by NSF CRII: CHS #1657599.

References

[1] Adobe Kuler. http://kuler.adobe.com/.
[2] M. Adnan, M. Just, and L. Baillie. Investigating time series visualisations to improve the user experience. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5444–5455. ACM, 2016.
[3] D. Albers, M. Correll, and M. Gleicher. Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the 32nd annual ACM conference on Human factors in Computing Systems*, pp. 551–560. ACM, 2014.
[4] 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.
[5] J. Bertin. *Semiology of Graphics*. University of Wisconsin Press, 1983.
[6] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.
[7] D. Brainard and B. Wandell. Asymmetric color matching: how color appearance depends on the illuminant. *Journal of the Optical Society of America A*, 9(9):1433–1448, 1992.
[8] C. A. Brewer. Prediction of simultaneous contrast between map colors with hunt's model of color appearance. *Color Research and Application*, 21(3):221–235, 1996.
[9] C. A. Brewer. Evaluation of a model for predicting simultaneous contrast on color maps. *The Professional Geographer*, 49(3):280–294, 1997.

[10] A. Brychtová and A. Çöltekin. The effect of spatial distance on the discriminability of colors in maps. *Cartography and Geographic Information Science*, pp. 1–17, 2016.
[11] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu. Facetatlas: Multifaceted visualization for rich text corpora. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1172–1181, 2010.
[12] R. C. Carter and L. D. Silverstein. Size matters: Improved color-difference estimation for small visual targets. *Journal of the Society for Information Display*, 18(1):17–28, 2010.
[13] W. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.
[14] M. Correll, D. Albers, S. Franconeri, and M. Gleicher. Comparing averages in time series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1095–1104. ACM, 2012.
[15] Ç. Demiralp, M. S. Bernstein, and J. Heer. Learning perceptual kernels for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1933–1942, 2014.
[16] M. Fairchild. *Color Appearance Models*. J. Wiley, 2005.
[17] C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):521–530, 2017.
[18] M. Harrower and C. Brewer. Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003. doi: 10.1179/000870403235002042
[19] C. G. Healey. Choosing effective colours for data visualization. In *Proceedings of Visualization'96*, pp. 263–270. IEEE, 1996.
[20] J. Heer and M. Stone. Color naming models for color selection, image editing and palette design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1007–1016. ACM, 2012.
[21] M. Jacomy. iWantHue. http://tools.medialab.sciences-po.fr/iwanthue/.
[22] S. Kaski, J. Venna, and T. Kohonen. Coloring that reveals high-dimensional structures in data. In *Neural Information Processing, 1999*, vol. 2, pp. 729–734. IEEE, 1999.
[23] P. Kovesi. Good colour maps: How to design them. *arXiv preprint arXiv:1509.03700*, 2015.
[24] J. Krantz. Stimulus delivery on the web: What can be presented when calibration isn't possible. *Dimensions of Internet Science*, pp. 113–130, 2001.
[25] C. I. D. L'Eclairage. Recommendations on uniform color spaces-color difference equations, psychomettic color terms. *Paris: CIE*, 1978.
[26] S. Lee, M. Sips, and H.-P. Seidel. Perceptually driven visibility optimization for categorical data visualization. *IEEE Transactions on visualization and computer graphics*, 19(10):1746–1757, 2013.
[27] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer. Selecting semantically-resonant colors for data visualization. In *Computer Graphics Forum*, vol. 32, pp. 401–410. Wiley Online Library, 2013.
[28] M. Livingston and J. W. Decker. Evaluation of trend localization with multi-variate visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2053–2062, 2011.
[29] M. R. Luo, G. Cui, and B. Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application*, 26(5):340–350, 2001.
[30] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2496–2505, 2012.
[31] M. Mahy, L. Van Eycken, and A. Oosterlinck. Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV. *Color Research & Application*, 19(2):105–121, 1994.
[32] R. McDonald and K. J. Smith. Cie94-a new colour-difference formula. *Coloration Technology*, 111(12):376–379, 1995.
[33] M. Millodot. *Dictionary of Optometry and Visual Science*. Elsevier Health Sciences, 2014.
[34] S. Mittelstädt, D. Jäckle, F. Stoffel, and D. A. Keim. Colorcat: Guided design of colormaps for combined analysis tasks. *Computer Graphics Forum*, 2015.
[35] S. Mittelstädt, A. Stoffel, and D. A. Keim. Methods for compensating contrast effects in information visualization. In *Computer Graphics Forum*, vol. 33, pp. 231–240. Wiley Online Library, 2014.
[36] N. Moroney, M. D. Fairchild, R. W. Hunt, C. Li, M. R. Luo, and T. Newman. The ciecam02 color appearance model. In *Color and Imaging Conference*, vol. 2002, pp. 23–27. Society for Imaging Science and Technology, 2002.
[37] N. Moroney and H. Zeng. Field trials of the CIECAM02 color appearance model. *Publications-Commission Internationale De L'Eclairage CIE*, 153:D8–2, 2003.
[38] K. Mullen. The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *The Journal of Physiology*, 359(1):381–400, 1985.
[39] B. Oicherman, M. Luo, B. Rigg, and A. Robertson. Effect of observer metamerism on colour matching of display and surface colours. *Color Research and Applications*, 33(5):346–359, 2008.
[40] L. Padilla, P. S. Quinan, M. Meyer, and S. H. Creem-Regehr. Evaluating the impact of binning 2d scalar fields. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):431–440, 2017.
[41] K. Reinecke, D. R. Flatla, and C. Brooks. Enabling designers to foresee which colors users cannot see. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2693–2704. ACM, 2016.
[42] P. Rizzo, A. Bierman, and M. Rea. Color and brightness discrimination of white leds. In *International Symposium on Optical Science and Technology*, pp. 235–246. International Society for Optics and Photonics, 2002.
[43] A. Robertson. Historical development of CIE recommended color difference equations. *Color Research and Applications*, 15(3):167–170, 2007.
[44] F. Samsel, M. Petersen, T. Geld, G. Abram, J. Wendelberger, and J. Ahrens. Colormaps that improve perception of high-resolution ocean data. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 703–710. ACM, 2015.
[45] A. Sarkar, L. Blondé, P. Le Callet, F. Autrusseau, P. Morvan, J. Stauder, et al. A color matching experiment using two displays: design considerations and pilot test results. In *Proceedings of the Fifth European Conference on Color in Graphics, Imaging and Vision*, 2010.
[46] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, 2017.
[47] S. Silva, B. Sousa Santos, and J. Madeira. Using color in visualization: A survey. *Computers & Graphics*, 35(2):320–333, 2011.
[48] A. Silverman, C. Gramazio, and K. Schloss. The dark is more (dark+) bias in colormap data visualizations with legends. *Journal of Vision*, 16(12):628–628, 2016.
[49] M. Stokes, M. Fairchild, and R. Berns. Precision requirements for digital color reproduction. *ACM Transactions on Computer Graphics*, 11(4):406–422, 1992.
[50] M. Stone, D. A. Szafir, and V. Setlur. An engineering model for color difference as a function of size. In *Color and Imaging Conference*, vol. 2014, pp. 253–258. Society for Imaging Science and Technology, 2014.
[51] D. A. Szafir, M. Stone, and M. Gleicher. Adapting color difference for design. In *Color and Imaging Conference*, vol. 2014, pp. 228–233. Society for Imaging Science and Technology, 2014.
[52] C. Tominski, G. Fuch, and H. Schumann. Task-driven color coding. In *Information Visualisation, 2008. IV'08. 12th International Conference*, pp. 373–380. IEEE, 2008.
[53] L. Wang, J. Giesen, K. T. McDonnell, P. Zolliker, and K. Mueller. Color design for illustrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1739–1754, 2008.
[54] M. Wijffelaars, R. Vliegen, J. J. Van Wijk, and E.-J. Van Der Linden. Generating color palettes using intuitive parameters. In *Computer Graphics Forum*, vol. 27, pp. 743–750. Wiley Online Library, 2008.
[55] L. Zhou and C. Hansen. A survey of colormaps in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(8), 2016.