Somewhere Over the Rainbow: An Empirical Assessment of Quantitative Colormaps

Yang Liu
University of Washington
Seattle, WA, USA
yliu0@cs.washington.edu

Jeffrey Heer
University of Washington
Seattle, WA, USA
jheer@uw.edu

ABSTRACT

An essential goal of quantitative color encoding is the accurate mapping of perceptual dimensions of color to the logical structure of data. Prior research identifies weaknesses of "rainbow" colormaps and advocates for ramping in luminance, while recent work contributes multi-hue colormaps generated using perceptually-uniform color models. We contribute a comparative analysis of different colormap types, with a focus on comparing single- and multi-hue schemes. We present a suite of experiments in which subjects perform relative distance judgments among color triplets drawn systematically from each of four single-hue and five multi-hue colormaps. We characterize speed and accuracy across each colormap, and identify conditions that degrade performance. We also find that a combination of perceptual color space and color naming measures more accurately predict user performance than either alone, though the overall accuracy is poor. Based on these results, we distill recommendations on how to design more effective color encodings for scalar data.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Colormaps; Color Models; Graphical Perception; Visualization; Quantitative Methods; Lab Study.

INTRODUCTION

The rainbow colormap – a scheme spanning the most saturated colors in the spectrum – has been a staple (or depending on one's perspective, eyesore) of visualization practice for many years. Despite its popularity, researchers have documented a number of deficiencies that may hinder accurate reading of the visualized data [4, 26, 36, 42]. They raise the following criticisms: the rainbow colormap is unfriendly to color-blind users [26], it lacks perceptual ordering [4], it fails to capture the nuances of variations for data with high spatial frequencies [36], and it is ineffective at conveying gradients due to banding effects at hue boundaries [4, 42].

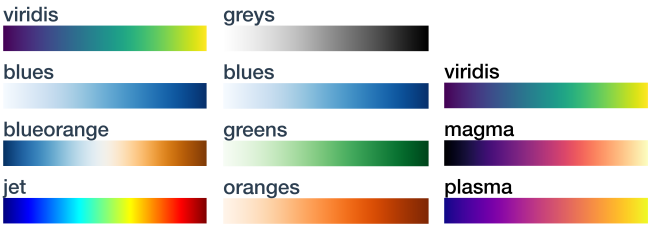

Figure 1: **Colormaps under study.** We evaluate four single-hue, three perceptually-uniform multi-hue, a diverging, and a rainbow colormap(s). We divide them into (a) assorted, (b) single-hue and (c) multi-hue groups, with two colormaps repeated across groups for replication.

Each of these problems may be traced to a naïve ramping through the space of color hues. In response, a common color design guideline for scalar quantitative data is to use a single-hue colormap that ramps primarily in luminance [6] (from dark to light, or vice versa). Changes in luminance provide a strong perceptual cue for ordering, consistent across individuals and cultures. Moreover, the human visual system has higher-resolution processing pathways for achromatic vision than for chromatic vision [23], supporting discrimination of higher spatial frequencies in the luminance channel.

These considerations raise a natural question: are the above criticisms specific to the rainbow colormap, or do they apply to multi-hue colormaps more generally? Defenders of rainbow and other multi-hue colormaps may cite not only aesthetic concerns, but also a potential for increased visual discrimination. By ramping through hue in addition to luminance, might viewers benefit from greater color separation across a colormap and thereby discern both small and large value differences more reliably? New multi-hue colormaps – the *viridis* colormap and its variants [38] – have recently been adopted by many visualization tools as a replacement for rainbow colormaps. These colormaps were formed by tracing curves through a perceptually-uniform color model, simultaneously ramping in both hue and luminance, while avoiding red-green contrast to respect the most common form of color vision deficiency.

Though existing guidelines and designs for quantitative color derive from both first principles and experience, they have not been comprehensively evaluated. In this work, we investigate

CHI 2018, April 21–26, 2018, Montréal, QC, Canada.

ACM ISBN 978-1-4503-5620-6/18/04 ...$15.00.
http://dx.doi.org/10.1145/3173574.3174172

the efficacy of a range of colormaps for encoding quantitative information. We examine a space of colormaps including a rainbow colormap, single-hue colormaps that vary primarily in luminance, multi-hue colormaps that vary both in hue and luminance, and (for comparison) a diverging colormap that combines opposing single-hue colormaps to convey distance from a neutral mid-point.

Our primary contribution is a comparative performance profile of quantitative color encodings. We analyze the speed and accuracy of each colormap in supporting relative similarity judgments across varying scale locations and value spans. We find that, when judiciously designed, single- and multi-hue colormaps both support accurate decoding. However, we find that single-hue colormaps exhibit higher error over small data value ranges, supporting the argument that multi-hue colormaps can provide improved resolution. In addition, we identify conditions that degrade accuracy across colormaps, notably that dark regions set against a white background afford much worse color discrimination than that predicted by perceptual color space models. We also confirm that a naïve rainbow colormap performs the worst among all colormaps considered. These results provide guidance for selecting effective quantitative colormaps and further improving their design.

As a secondary contribution, we construct statistical models to predict user performance on triplet comparions tasks, based on color theory measures. We consider both perceptual color spaces such as CIE LAB [24] and CAM02-UCS [28], as well as a model of color naming [21]. We find that combining perceptual measures with color naming measures leads to higher predictive accuracy than either alone. However, we also observe that our models fail to account for a large proportion of the variance observed in our experiments, suggesting the need for future work on refined color measures applicable to automated design and evaluation.

RELATED WORK

We draw on both the century-long research on color theory, and more recent work on colormap design and evaluation.

Color Models

Perceptually-uniform color spaces attempt to model equal perceptual differences as equal distances in a vector space [43]. The color science community has progressively refined a series of models for improved estimation accuracy over a wider variety of viewing conditions. Example models include CIELAB [24], ΔE_{94} [30], DE2000 [29], and CAM02-UCS [28].

Despite being one of the earliest perceptually uniform models, CIELAB remains a popular choice in visualization research (e.g. [25, 40]), thanks to its relatively simple color distance calculation equation, which is the L^2 Euclidean norm between two points in the space. CAM02-UCS is a recent variant that builds upon the CIECAM02 color appearance model and provides better estimation of lightness and hue. Dozens of empirical datasets, which contain pairs of color difference values with an average of 10 ΔE^*_{ab} units, were employed in the development of the CAM02-UCS model. In this paper, we use CIELAB and CAM02-UCS for our analyses. We use the LAB implementation of D3 [5], which assumes a D65 standard illuminant as the white point. For CAM02-UCS, we use Connor Gramazio's JavaScript implementation [16].

While uniform color models offer useful approximations of perceived color difference, they omit factors that may influence color perception. Properties of the color stimuli, such as the size of the color patches [10, 39], the spatial distance between two colors [9], and the geometric mark types [40] can modulate color discriminability. In addition, the surrounding context in which the color is presented can result in large distortions of color perception due to simultaneous contrast [7, 11, 42]. Even when model predictions rigorously align with perceived differences, color distance models do not account for visual aesthetic experiences as in color harmony [11] and aesthetic preference [32] theories. Demographics and color vision variations of the viewers may also affect our ability to discriminate colors [34]. In our experiments, where possible we seek to control factors that may interfere with color perception, but we acknowledge we have limited environmental control given our use of crowdsourcing platforms.

In addition to perceptual modeling efforts, psychologists have investigated the extent to which the linguistic labels assigned to colors shape our perception (see Regier & Kay [33] for a survey). A number of controlled experiments find that color naming can affect categorization and memory. For example, Russian speakers may more quickly discriminate two different shades of blue, as the Russian vocabulary contains two basic color terms for blue [45].

To quantify the association of names to color, researchers have proposed various models. Chuang et al. [12] formulate a non-parametric probabilistic model and introduce a measure of name saliency based on the entropy of the probability distribution. Heer & Stone [21] extend this model to introduce similarity metrics of color names, and contribute a mapping between colors and names by applying the model to a large web survey containing over 3 million responses. We use their model in our analyses of color naming in this paper. These models provide measures to quantitatively analyze categorical perception effects due to color names.

Colormap Design & Evaluation

As color is an important visual channel in visualization, the design of appropriate colormaps has received much attention (see [37] or [46] for surveys). Predefined colormaps are developed based on perceptual and cognitive heuristics, designer experience, application of color models, empirical data from experiments, or a combination thereof. For example, the ColorBrewer [18] schemes are informed by color theory, with the final colors hand-tuned for aesthetic appearance. The design of the *viridis* [38] colormap focuses on perceptual uniformity, ramping in both hue and luminance through equal steps in the CAM02-UCS color space.

A number of interactive systems and algorithms also exist to aid users in constructing or selecting color schemes. The early PRAVDA system [3] takes into consideration data types, anticipated tasks, and perceptual properties when recommending appropriate colormaps. Subsequent research focuses on perceptual saliency [25], separation [19], semantic resonance

of color/category associations [27], visual aesthetics [41] and energy consumption of display devices [13]. Colorgorical [17] combines the scores of perceptual distances, color names, and aesthetic models to automatically generate categorical palettes.

Prior work has also sought to empirically evaluate univariate quantitative color encodings [7, 20, 31]. Ware [42] conducts multiple experiments to evaluate (1) how accurately do people extract metric information from color encodings and (2) how well do colormaps preserve the form, or gradient of the underlying data. A recent work by Ware et al. [44] compares six colormaps, testing the highest visible spatial frequencies at varying locations. Brewer et al. [8] evaluate eight discrete schemes in supporting visualization tasks on choropleth maps. While we also provide a comparative analysis of quantitative colormaps, we instead focus on comparing single- and multi-hue colormaps in supporting similarity judgments. The "Which Blair Project" [35] develops an interesting perceptual method to evaluate luminance monotonicity of colormaps, which relies on our ability to distinguish human faces. Kindlmann et al. [22] further extend the idea to propose a technique for luminance matching. These two studies focus on luminance; here we are interested in assessing judgment performance across both hue and luminance.

EXPERIMENTAL METHODS

Our objective was to assess the effectiveness of each colormap for encoding scalar information. As prior work establishes that color is a poor visual channel for precise magnitude estimation [14], we are less interested in how well people extract the exact metric quantity from the colormap. Instead, we focus on ordinal judgments of relative difference: given a reference data point, how well can people judge which other points are most similar? We carried out a suite of three experiments to compare the perception of relative distances encoded by colormaps. Each experiment focused on a subset of colormaps in a within-subjects design; we ran a separate experiment for each group of colormaps in order to mitigate fatigue effects. To check the robustness of our results, we replicated two colormaps across groups. The general methods of each experiment are identical.

Task

Our experiments used an ordinal triplet judgment task: given a reference color and two alternative stimuli sampled on either side of the reference, participants judged which of the response stimuli is closest in distance to the reference. We selected this task for multiple reasons. First, compared to direct value estimation, a binary forced-choice response shifts the emphasis to more rapid, perceptual judgments. We are less interested in value estimation because other visual channels, such as position and length, are far superior than color in this task [14]. For example, viewers of a choropleth map of employment data likely spend more time comparing colored regions than they do resolving these to absolute values, answering questions such as "which U.S. state has a rate most similar to Michigan: Wisconsin or Ohio?" Second, compared to a simpler pair-wise ordinal task (i.e., participants see two stimuli and judge which represents a larger value), triplets allow us to assess *distance*, not just *ranking* relationships. Triplet judgments are more difficult than simple rank-order judgments, and so more likely to reveal discrepancies in colormap performance.

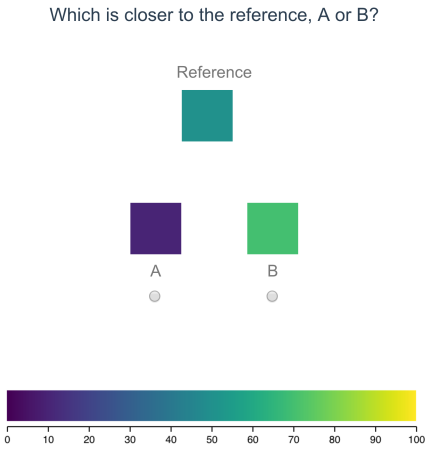

Figure 2: **Experiment interface.** Participants see a reference stimulus along with two choices, and pick which of these alternatives is closer in distance to the reference.

A color legend was included for reference in each presentation. We supplied the legend because legends influence color judgments in real world visualization tasks, potentially with conflicts between what one perceives with the colors alone and what one effortfully "reads" from the legend.

Stimuli

We included four single-hue and five multi-hue colormaps in our studies, grouped into the three sets shown in Figure 1. We use the term single-hue to denote colormaps varying primarily in luminance. Due to hand-tuning, the ColorBrewer [18] sequential colormaps we chose have subtle variations in hue, with the exception of *greys*. The first group (**assorted colormaps**) aimed to compare representative colormaps from four distinct types, following an extended version of Brewer's taxonomy [6]. We picked *viridis* from the multi-hue sequential type, *blues* from the single-hue sequential type, *blueorange* from the diverging type, and *jet* – long the default in MATLAB – to represent rainbow colormaps. The other groups focus on single-hue and multi-hue sequential variants. The second group (**single-hue colormaps**) includes *greys* (a baseline condition with purely achromatic shades), along with *blues*, *greens*, and *oranges*, three hues that occupy relatively opposing regions of LAB space. The third group (**multi-hue UCS colormaps**) includes multi-hue colormaps created using the UCS color space: *viridis*, *magma*, and *plasma*.

We rendered each visual stimulus as a 50×50 pixel color square against a white background. Admittedly, placing large color patches on a uniform background differs from many real-world heatmaps, and one might see additional effects in scalar field contexts (e.g., due to gradients). For this study, we chose to stay closer to the conditions for which the underlying color models are defined, contributing an actionable baseline for comparing colormaps and a comparison point for future studies. We controlled the size of the color patches, the background color, and the spatial layout of the stimuli to mitigate

potential confounds with mark size, simultaneous contrast, and spatial distance [9, 39]. We focused on white backgrounds as they are most common in both print and on-screen.

We generated the trial stimuli for each colormap in the following way. Assuming a data domain of $[0, 100]$, we first sampled reference points along uniform data value steps of 10 units along the color scale. For each reference point, we then sampled comparison values: one of lower value than the reference, and one higher. In each trial, one of these points is systematically farther away than the other.

We generated comparison points offset from the reference point using *spans* (total difference between highest and lowest point) of 15, 30, and 60. We included two trials for each combination of reference and span: one in which the lower value is nearer the reference, and vice versa. As a concrete example, for a reference of 50 and span 60 the sampled triplets are (30, 50, 90) and (10, 50, 70). To encourage a similar difficulty across spans, we adopted the logic of the Weber-Fechner Law [15], which predicts that the perceived change is in constant ratio to the intensity of the initial stimuli. In our case, we placed the more distant response stimulus at twice the distance (in data units) from the nearer. Pilot studies confirmed that this choice resulted in reasonable yet suitably difficult tasks; an earlier iteration with an offset half this size resulted in roughly double the error rate.

After generating all triplets, we discarded reference/span combinations with values outside the $[0, 100]$ domain. This resulted in too few trials in the span 60 condition, so we introduced two additional reference values (45, 55) for this span level only. This procedure produced 42 trials per colormap.

Participants

We recruited subjects via Amazon's Mechanical Turk (MTurk) crowdsourcing platform. Prior research has established the validity of crowdsourcing experiments for controlled quantitative modeling in color perception [34, 40]. While we sacrificed control over monitor display and situational lightning conditions, we gained samples from a wider variety of display conditions in the real-world web user population. In addition, the variance introduced by viewing conditions is partly accounted for by per-subject random terms in our statistical models. Each experiment run was implemented as a single Human Intelligent Task (HIT) to ensure a within-subjects design. We restricted the participants to be within the United States and to have an acceptance rate over 95%.

Procedure

We first screened the participants for color vision deficiencies using four Ishihara plates. As factors including uncalibrated displays and image compression can make Ishihara plates unreliable, we also stated in the consent page that participants must have normal color vision. The participants then read a tutorial page with a sample question, which encouraged them to use the color legend, explaining that the correct answer should be deduced from value differences in the legend. Prior to the experiment, we administrated a practice session consisting of 5 trials from an irrelevant colormap to reduce learning effects.

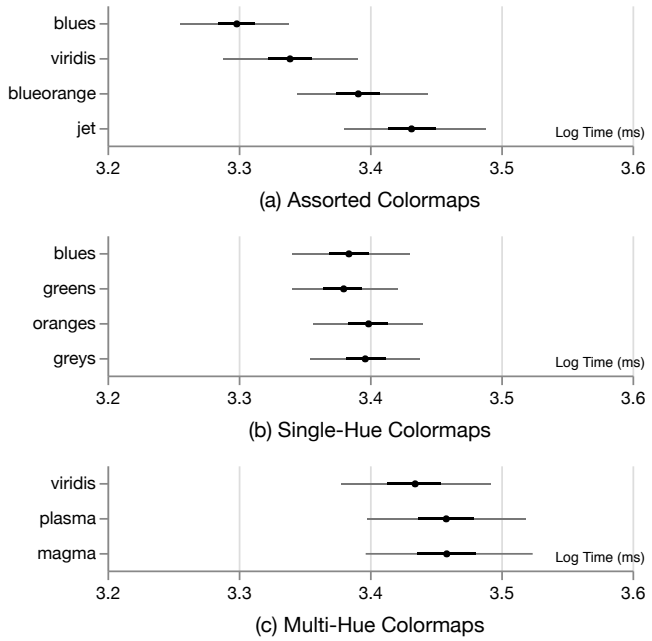

Figure 3: **Log response time by colormap for each study.** Plots depict bootstrapped means, with 50% (thick) and 95% (thin) CIs. (a) Assorted colormaps. The single-hue colormap *blues* is the fastest, followed by *viridis*. The rainbow colormap *jet* is the slowest. (b) Single-hue colormaps. Subjects spent almost identical time on average on each colormap. (c) Multi-hue colormaps. UCS multi-hue colormaps are comparable in speed. *Viridis* is slightly faster, but not significantly so.

Participants completed blocks of trials for each colormap, with an option to take breaks between sessions to mitigate fatigue. We asked subjects to respond as quickly and accurately as possible, prioritizing accuracy. We counterbalanced the colormap order using either a Balanced Latin Square or a full permutation of all possible orders, depending on the total number of colormaps in each study. We randomized the question order for each colormap. An engagement check question appeared randomly per colormap block to ensure attentive participation.

In each trial, we simultaneously presented the three color stimuli arranged in a triad, with a legend that included ticks at each 10 unit interval (Figure 2). Participants responded by clicking on the choice square and clicking the "Next" button, or by pressing the "a" or "b" key followed by "enter".

Data Analysis

Our primary dependent variables are log-transformed response time (RT) and an error label, indicating whether a subject answered the question correctly. Observing that RT follows a log-normal distribution, we performed log transformation. The error response uses a binary coding of 1:error, 0:correct. To visualize effect sizes, we calculate bootstrapped confidence intervals (created by sampling entire subjects, not individual responses, with replacement) and plot both 50% and 95% CIs.

Previous quantitative modeling on color perception has fit linear models to the mean proportion of response, obtained by averaging individual binary outcomes per cell [39, 40]. This

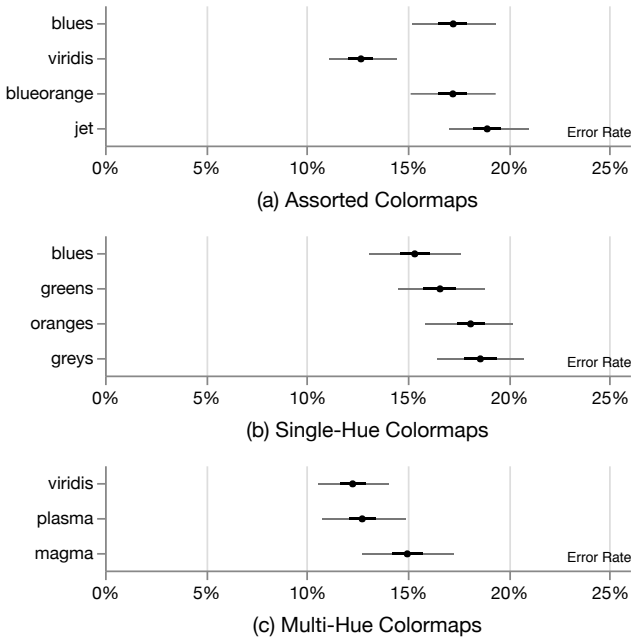

Figure 4: **Error rate by colormap for each study.** Plots depict bootstrapped means, with 50% (thick) and 95% (thin) CIs. (a) Assorted colormaps. *Viridis* excels in accuracy while *jet* is the most error-prone. (b) Single-hue colormaps. Though slightly faster, *blues* and *greens* have overlapping confidence intervals with the slower colormaps, *oranges* and *greys*. (c) Multi-hue colormaps. Multi-hue colormaps have comparable accuracy within group. The per-colormap average error rate of *magma* is higher as it contains degenerate cases.

approach discards a large portion of the individual variance. As a result, the fitted model describes the mean performance from a sample group of the population, but not the performance of any individual.

In this paper, we instead fit models to individual observations, using linear mixed-effects models for RT and logistic mixed-effects models for error (using the lme4 package in R [2]). Mixed-effects models can incorporate random effect terms to account for variation arising from subjects as well as other sources. In our models we include fixed effect terms for colormap, span, and their interaction. Following Barr et al. [1], we also include maximal random effects structures with per-subject terms for random intercept (capturing overall bias) and random slopes for each fixed effect (capturing varied sensitivities to experiment conditions). As we later show, specific colors may exhibit outlying performance relative to a colormap as a whole. In response, we include random intercepts for each unique reference color (i.e., colormap / reference value pair) to improve generalization of fixed effect estimates.

EXPERIMENTAL RESULTS

We now present the results from our three experimental runs. We first share the results from each colormap group, and then investigate special cases with surprisingly low or high error rates. Figures 3 and 4 show global time and error estimates per colormap. Figures 5, 6, 7, and 8 provide more detailed plots across span and reference conditions.

Across colormap groups we conducted a diagnostic analysis before examining time and error separately. In all cases we note a similar, positive correlation between response time and error: on average, subjects spend more time on the more difficult cases. This result suggests that the performance measures are not simply the result of varied speed/accuracy trade-offs.

Assorted Colormaps

A total of 56 subjects (19 female, 36 male, 1 other, $\mu_{age} = 35.3$ years, $\sigma_{age} = 8.9$ years) participated in the assorted colormap study. Subjects completed the study in 15 minutes on average and were compensated $2.00 USD.

Time: Blues & Viridis are Faster than BlueOrange & Jet

Likelihood ratio tests of linear mixed-effects models for log response time found significant main effects for colormap ($\chi^2(9) = 60.5$, $p < 0.001$), span ($\chi^2(8) = 60.0$, $p < 0.001$), and their interaction ($\chi^2(6) = 26.3$, $p < 0.001$). To compare response times across colormaps, we applied post-hoc tests with Holm's sequential Bonferroni correction. We find that both *blues* and *viridis* are significantly faster than *blueorange* ($p < 0.01$, both cases) and *jet* ($p < 0.001$, both cases). The difference in means between *blues* and *viridis* is not significant, nor is the difference between *blueorange* and *jet*.

With respect to span, subjects performed significantly slower when the span was 60 compared to a span of 30 ($p < 0.01$) or 15 ($p < 0.05$). The significant interaction between colormap and span stems primarily from *blues*, which was relatively slow for small spans. As we will discuss shortly, this decrease in performance correlates with more pronounced errors.

Subjects made faster judgments with the *viridis* and *blues* colormaps and spent more time determining distances with *blueorange* and *jet*, presumably because the distances are not as apparent. This discrepancy may result from increased effort discerning perceptual similarities and/or consulting color legends. Across all colormaps, more time was needed when colors were further apart in the color scale.

Error: Viridis Excels; Blues Degrades for Low Spans

Tests of logistic mixed-effects models for error again found significant effects of colormap ($\chi^2(9) = 46.0$, $p < 0.001$), span ($\chi^2(8) = 42.9$, $p < 0.001$), and their interaction ($\chi^2(6) = 28.6$, $p < 0.001$). Post-hoc tests revealed that *viridis* is less error-prone than *blues* and *jet* (both $p < 0.001$). Across colormaps participants made fewer mistakes on average in the smallest span compared to other levels (both $p < 0.001$). The interaction effect again stems from the differential characteristics of *blues*: when the span was small, error increased. An example of such triplets is (20, 30, 35).

In a follow-up analysis where for all colormaps we dropped responses for span 15, a significant effect of colormap on error rate ($p < 0.001$) remains, but without a significant interaction. In this case we did not observe a significant difference between *viridis* and *blues* in error rate, but *blues* outperforms *jet* and *blueorange* ($p < 0.05$).

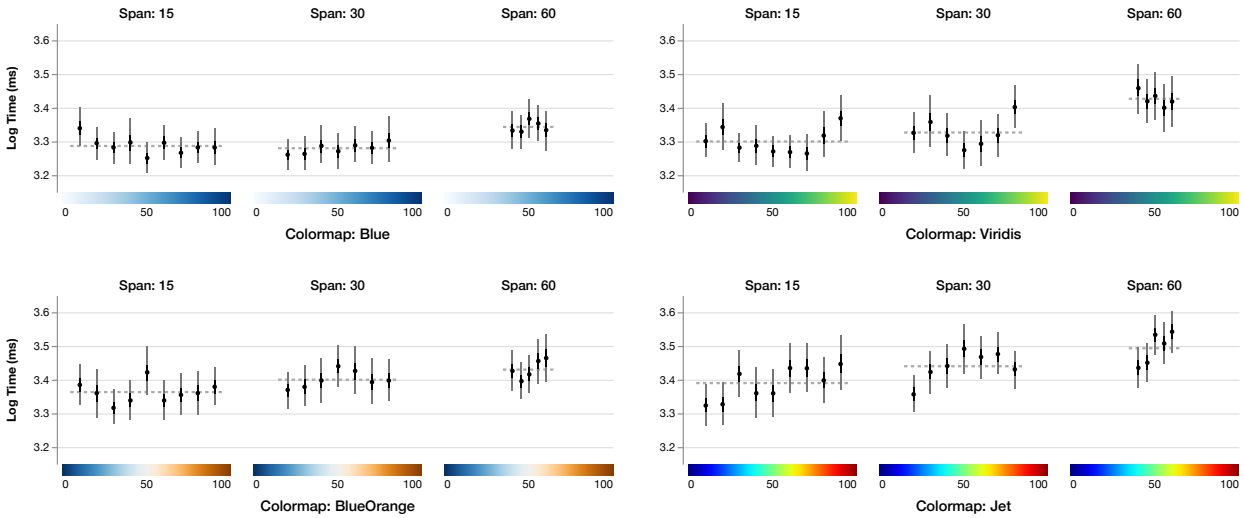

Figure 5: **Log response times by span and reference for assorted colormaps.** Points indicate bootstrapped means, along with 50% (thick) and 95% (thin) CIs. Each sub-plot includes the mean value for each span level (dotted grey line). Across colormaps, response times increase with larger spans. *Jet* exhibits the longest response times.

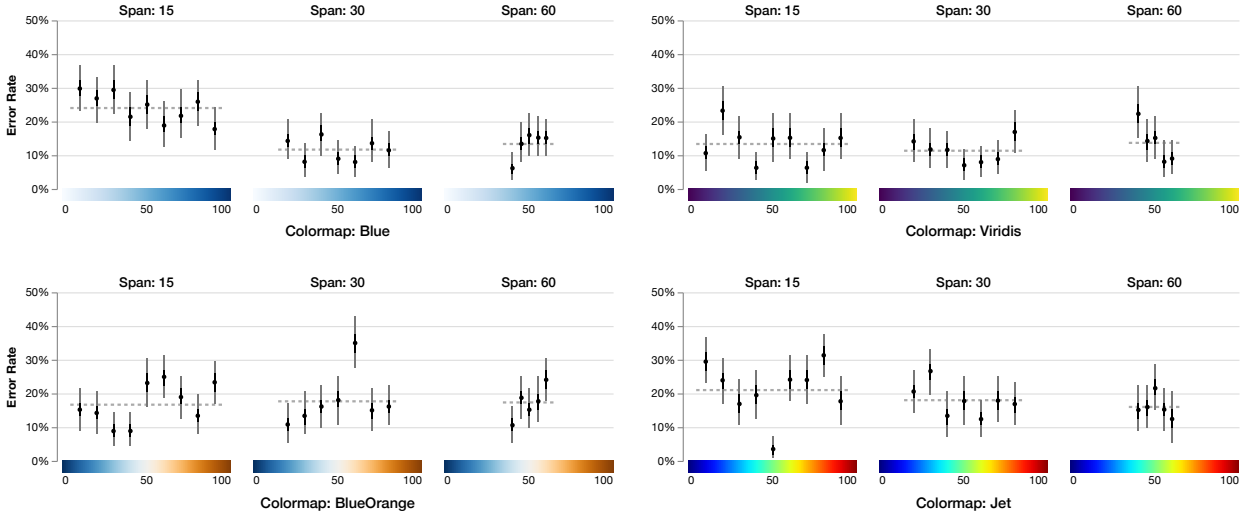

Figure 6: **Error rate by span and reference for assorted colormaps.** Points indicate bootstrapped means, along with 50% (thick) and 95% (thin) CIs. Each sub-plot includes the mean value for each span level (dotted grey line). *Viridis* exhibits consistently low error across the board. The accuracy of *blues* matches that of *viridis* at larger spans, but drops notably for the smallest span. The *blueorange* diverging scheme exhibits errors when comparison is made across the central blue-orange hue boundary.

Summary

In this study, *viridis* demonstrated both superior speed and accuracy. *Blues* performed comparably well at spans 30 and 60: it was fast and accurate so as long as there was sufficient spacing between adjacent colors. However, once the colors were too close in the color scale, the accuracy of *blues* dropped considerably, together with a mild increase in response time. The diverging colormap *blueorange* and the rainbow colormap *jet* were both slower and more error-prone. We examine special cases affecting these latter two colormaps later in the paper.

Comparing with the subsequent studies, we note similar error results for replicated colormaps, but systematically lower response times in the assorted colormaps group (Figure 3). We attribute this disparity in part to individual differences. For example, 64.3% of participants were male in the assorted group, while single-hue and multi-hue groups were 33.9% and 42.6% male respectively. In a linear mixed-effects model of RT with gender as the fixed effect, fit to data from all three experiments, the male group was significantly faster ($p < 0.01$). We found no significant effect of gender in a similar model for error.

Single-Hue Colormaps

56 subjects (36 female, 19 male, 1 other, $\mu_{age} = 37.2$, $\sigma_{age} = 11.1$) were assigned single-hue colormaps. Subjects averaged 15 minute sessions and were compensated $1.60 USD.

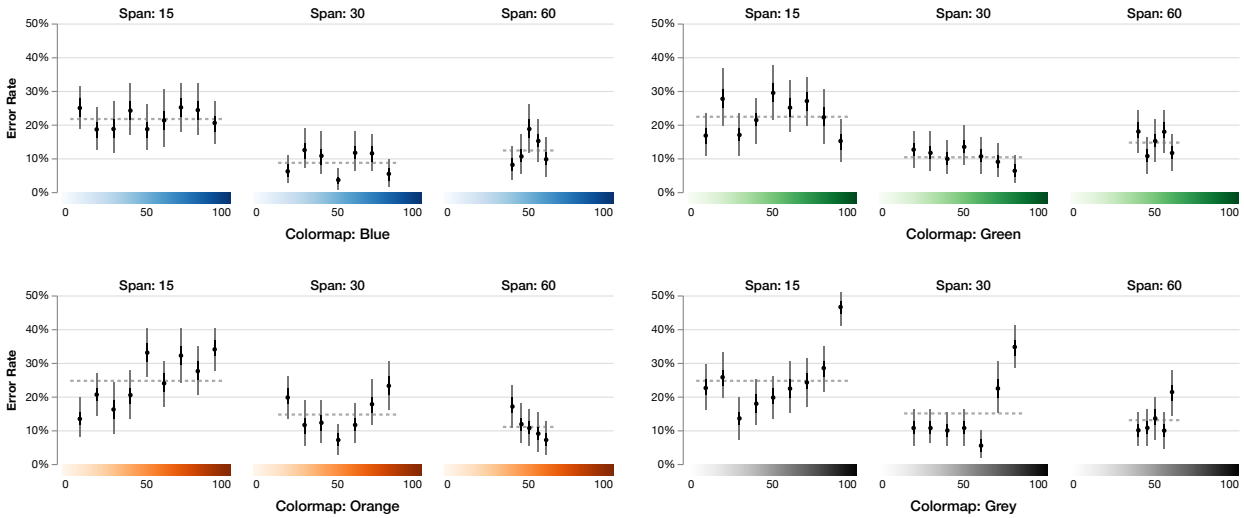

Figure 7: **Error rate by span and reference for single-hue colormaps.** Points indicate bootstrapped means, along with 50% (thick) and 95% (thin) CIs. Each sub-plot includes the mean value for each span level (dotted grey line). All single-hue colormaps similarly suffer from resolution issues when the span is small. *Greys* degenerates in low luminance regions.

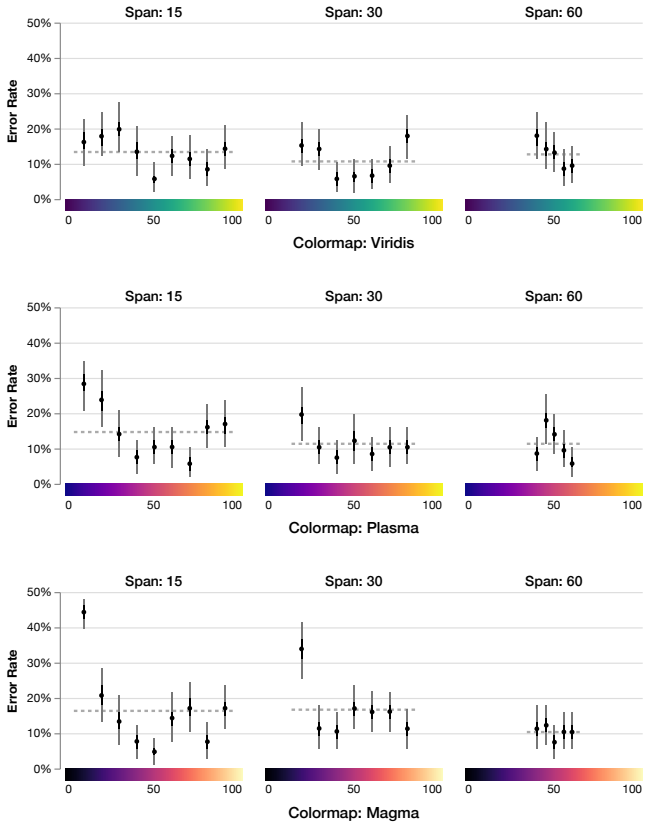

Figure 8: **Error rate by span and reference for multi-hue UCS colormaps.** Points indicate bootstrapped means, with 50% (thick) and 95% (thin) CIs. Each sub-plot includes the mean value for each span level (dotted grey line). We observe similar dynamics across colormaps. Performance degrades in the dark region of *magma*, and to a lesser extent of *plasma*.

Time: No Differences in Single-Hue Responses

In a linear mixed-effects model of RT, we found a significant effect of span ($\chi^2(8) = 39.9$, $p < 0.001$), but not for colormap or their interaction. This result is consistent with the per-colormap means plotted in Figure 3: participants have similar response times ($\sim 10^{3.39} = 2,454$ milliseconds) for each colormap in the single-hue group.

Error: Single-Hue Colormaps Suffer from Low Resolution

Tests with a logistic mixed-effects model of error found a significant effect of span ($\chi^2(8) = 86.0$, $p < 0.001$), but no significant colormap or interaction effect. This result is consistent with Figure 4: despite lower means, 95% CIs for *blues* and *greens* overlap with those for *oranges* and *greys*. Looking across studies, we see very similar error profiles for *blues* in Figures 6 and 7, indicating successful replication.

Post-hoc comparisons confirmed that error rates for span 15 are significantly higher than span 30 ($p < 0.05$) or span 60 ($p < 0.05$). This result corroborates the increased errors for *blues* in low-span cases in the prior study, and extends it to a larger family of single-hue colormaps. These colormaps suffer from poor resolution for nearer value comparisons.

To further test this claim, we calculated the LAB distances between the reference stimulus and the two choices respectively, and subtracted them to obtain a *difference* measure in units of ΔE. We found that in low-span conditions where accuracy plummets, the ΔE difference is around 5, close to the just-noticeable difference (JND) found in practical situations [39, 40]. Though the ΔE between each stimulus is large enough for the colors to be distinguishable, the difference in ΔE between pairs is hard to discriminate, leading to increased error.

Multi-Hue UCS Colormaps

54 subjects (31 female, 23 male, $\mu_{age} = 36.7$, $\sigma_{age} = 10.1$) participated in the multi-hue colormap study. We discarded

data from 1 subject (2%) due to missing responses. Subjects averaged 12 minute sessions and paid $1.20 USD.

Time & Error: Multi-Hue Colormaps are Comparable

Analysis of the multi-hue UCS colormaps detected no significant differences between colormaps in terms of either response time or error rate. Figure 3 shows that the mean response times align around 2.82 seconds ($10^{3.45}$ milliseconds). Similarly, the mean error rates of *viridis* and *plasma* are slightly lower than that of *magma*, but exhibit overlapping 95% CIs (Figure 4). The more detailed plots in Figure 8 reveal spikes in error rate for *magma*, and to a lesser extent *plasma*, around low reference values. We examine this issue further in the next section.

Multi-Hue UCS Colormaps have Lowest Error Across Studies

Comparing across studies, the error profiles for *viridis* in Figures 6 and 8 are quite similar, indicating successful replication. We see that across studies the UCS colormaps exhibit the lowest error rates, though with slightly longer response times.

Analysis of Special Cases

The above section analyzes colormaps in terms of their mean performance, with models that include random effects to account for some of the larger swings among specific reference points. Here we perform a complementary analysis, investigating the specific conditions in which error rates are surprisingly high or low. We take a closer look at (1) error increases in low luminance conditions (*greys*, *magma*, *plasma*), (2) the performance of the diverging *blueorange* colormap, and (3) a special case where *jet* – the colormap with worst performance overall – exhibits extremely low error.

Performance Degrades in Low Luminance Regions

An obvious abnormality across studies and colormaps is a dramatic increase of error rates in the black regions, particularly *greys*, *magma*, and, to a lesser degree, *plasma* (Figures 7 and 8). For example, the *magma* triplets ■■■ (0, 10, 15) and ■■■ (5, 10, 20) exhibit high error. The affected conditions all involve small values in the luminance channel; the low luminance level appears to afford much worse color discrimination than that predicted by either the LAB or UCS perceptual models. This observation is likely specific to our choice of a white background, with the high contrast impeding the discrimination of dark shades. We hypothesize that an analogous shortcoming will occur for high luminance shades set against a dark background.

BlueOrange Suffers when Values Straddle the Mid-Point

A closer look at *blueorange* suggests a primary source of errors (Figure 6). When all three triplet colors lie on a single-hue half of *blueorange*, the performance closely matches that of the corresponding single-hue colormap. For example, the first three points in the small span plot of *blueorange* average about 10%, similar to the mean error rates of *blues* for the medium span (Figure 6, 7). Note that we double the span to compare to single-hue colormaps, as each hue takes up one-half of the range of the diverging colormap. As indicated by the high error rates in the middle of *blueorange*, subjects were prone to mistakes when making comparisons across the blue-orange boundary. A representative triplet is ■■■ (50, 60, 80), where the lower, achromatic option is closer than the similarly-hued, but much more saturated, option. This result suggests that diverging colormaps may be less accurate in situations involving comparisons with the mid-point, perhaps due to erroneous grouping of chromatic colors versus a nearer achromatic color.

Where the Rainbow Shines: Color Name Association

Though the majority of reference stimuli in *jet* lead to higher error than other colormaps, reference value 50 performs remarkably well at span 15 (Figure 6). In the small span condition this reference point has a mean error rate as low as 3.5%, which is among the lowest in all observations! The corresponding color triplets are ■■■ (40, 50, 55) and ■■■ (45, 50, 60). These triplets lie in an isoluminant region of *jet*: there are no luminance cues that might suggest ordering. Instead, these triplets happen to straddle color name boundaries that align with the underlying value differences. Color name distances [21] from the reference average 0.23 and 0.94 for the nearer and further values, respectively. The first triplet has modal names of `cyan` versus two `green`s, while the second triplet has two `green`s versus `yellow`. This result suggests that categorical effects, or *banding* by name, can contribute to improved discrimination if applied in the right direction and, conversely, may hamper perception if dischordant with the true value difference.

COLOR MODEL ANALYSIS

In addition to empirical characterization of user performance, we would like to have a theoretical model. For example, given a previously untested colormap, might we predict its relative performance? If so, we could use the model to automatically optimize colormap designs. To assess this question we construct a series of models that attempt to generalize beyond the specific colormaps using a set of three color distance models:

- **LAB**: The CIELAB color space [24].
- **UCS**: The CAM02-UCS uniform color space model [28].
- **Name**: The color name model of Heer & Stone [21].

The first two color models (LAB and UCS) provide *perceptual* color spaces that approximately model perceptually uniform color distances. We include both for comparison. For LAB, we use Euclidean distance (ΔE) to measure color distance. The third model (Name) is a model of *categorical* effects that measures color difference by comparing the distributions of observed color terms (e.g., orange, blue, fuchsia) that people use to label color swatches. The Name model is included to capture categorical effects of color naming that may not be reflected by the perceptual models. Following prior work [21], we use a cosine distance measure between color term vectors.

To apply these measures to a triplet comparison task, we first compute the color model differences between the reference stimulus and the two response stimuli. We then calculate the difference of the predicted color model distances; i.e., we simply subtract the distance value for the correct answer from the distance value for the incorrect answer. A negative difference indicates that the correct answer (the more similar data value) is further away according to the distance measure. A positive difference indicates a larger distance for the incorrect answer (the more dissimilar data value).

Model	df	AIC	BIC	logLik	deviance
LAB	24	21668	21863	-10810	21619
UCS	24	21665	21860	-10808	21617
Name	24	21585	21781	-10769	21537
UCS + Name	63	21308	21821	-10591	21182
UCS * Name	288	21377	23723	-10401	20801

Table 1: Diagnostics for error models based on color model distances. Columns indicate degrees of freedom (df), AIC and BIC model selection scores, log-likelihood (logLik) and deviance. An additive model with UCS and color name difference terms achieves the best balance of fit and parsimony according to AIC and BIC scores (lower is better).

Parameter	Estimate	Std. Error	P-Value	
Intercept	-1.0848	0.1804	< 0.001	***
UCS_Q2	-0.4031	0.2043	0.048	*
UCS_Q3	-0.5298	0.1618	0.001	**
UCS_Q4	-0.4452	0.2482	0.073	.
Name_Q2	-0.5009	0.1641	0.002	**
Name_Q3	-0.6309	0.1621	< 0.001	***
Name_Q4	-0.6207	0.1336	< 0.001	***

Table 2: Fixed effect parameter estimates and p-values for a logistic regression model (UCS + Name) of judgment error. Increasing UCS and Name difference lead to lower error, but this effect attenuates in the highest quartile.

Error Analysis

To predict error rates, we fit a logistic regression model. We use mixed-effects models with random effect terms for both subject (to account for variance due to individual differences) and colormap (each trial includes presentation of a color legend, and we account for this in order to estimate more generalizable fixed effects). We use maximal random effects structures [1], with intercepts for each random effect and corresponding random slope terms for each fixed effect.

We first assessed which form the predictor should take. We examined both direct use of color model difference estimates (a continuous, linear predictor) and binned factors based on quartile boundaries (a discrete, potentially non-linear predictor). All fitted models exhibit statistically significant fixed effect estimates, via both Wald z-tests and Likelihood Ratio tests. The binned predictor leads to better models for all color difference types: with improved fit (log-likelihood and deviance) and lower model selection scores (AIC, BIC). As a result, we focus on the discrete predictors.

Next, we compare these single-effect models to assess performance differences among color difference types. Which color model most accurately predicts performance? Table 1 shows the resulting model diagnostics. We see that name difference performs the best according to all measures. The UCS model outperforms LAB, but by a miniscule margin. Overall, the differences between the three models are small.

We then fitted two-factor models that include perceptual and categorical terms. For the perceptual term we chose UCS rather than LAB for two reasons. First, UCS performs slightly better than LAB as a single predictor. Second, the color name model internally applies a fine-grained discretization of the LAB color space, and so is likely to exhibit higher correlation with LAB. We built models both with and without interaction terms. The last two rows of Table 1 show the resulting model diagnostics. Both models improve upon the single-factor models in terms of fit and AIC score. The model with interaction terms exhibits improved fit (higher log-likelihood and lower deviance), but this is unsurprising given the greater degrees of freedom. The additive model has lower AIC and BIC scores than the full model, indicating a more parsimonious model. To avoid overfitting, we stop with the additive model.

Table 2 shows the coefficients of the resulting model. The intercept term is the logit value for triplets with difference values residing in the first quartiles for both UCS and Name. As the color differences increase, we see increasingly negative coefficients, indicating lower error rates. However, for both UCS and Name this trend tapers off for the highest quartile (Q4): relative to the earlier quartile (Q3), the error slightly increases for the largest color differences. This effect may stem from issues with large distances in perceptual color spaces: perceptually uniform color spaces were constructed in accordance with empirical color discrimination judgments at a small scale (e.g., 10-20 ΔE [28]). As a result, longer scale distances in these models are known to be more inaccurate.

How well do these color models predict user performance overall? To assess this question, we can use the additive model to predict the average performance across all experimental conditions. While this is "testing on the training data" and so not a means of assessing generalization, it nevertheless serves as a useful diagnostic. Comparing the model's predicted error rates with the observed rates via standard linear regression, we achieve of an R^2 value of 0.108. In other words, our fitted model only explains about 10% of the observed variance.

We can also examine model predictions for the average performance of each colormap: does our model rank the colormaps in an order similar to the observed error rates? The Spearman rank correlation between the model predictions and the observed empirical error rates ($\rho = 0.45$) is not high and not statistically significant. In short, the fitted model does an unsatisfactory job of predicting overall colormap performance.

Time Analysis

To analyze timing responses, we followed a similar procedure as we did for the error analysis, but using linear mixed-effects models of the log-transformed response times rather than logistic regression. Once again, the binned variants outperform the linear predictors. For the single-factor models, UCS outperforms LAB, which outperforms Name. Comparing a full model with UCS, Name, and interaction terms to a model without an interaction term again finds that the full model exhibits worse AIC and BIC scores.

Using the additive (UCS + Name) model to predict per-condition average response times in the log domain results in

an R^2 value of 0.244, accounting for 24% of the observed variance. The rank correlation of observed per-colormap average responses with model predictions ($\rho = 0.67$) is higher than for error, but again is not statistically significant.

Summary

Combining perceptual color models and color naming models leads to higher predictive accuracy for both time and error than either alone. This suggests that lower-level perception and language-level processes may both play a role in the interpretation of quantitative color encodings. We also observe that increasing perceptual and name differences correlate with higher judgment accuracy, but that this trend is non-linear, tapering off among the highest quartile of differences for both measures. That said, we believe the primary take-away is a need for caution, as neither the error model nor time model lead to accurate prediction of the observed experimental results (let alone for new, unseen conditions).

Improved models or measures could lead to more accurate predictions of user performance. Some issues may arise from the triplet comparison task: perceptual color models are fit to pairwise discrimination judgments, and so may be less well-suited for the comparison tasks studied here. Moreover, our measures of difference do not take into account either the relative color space locations or the magnitude of the underlying color distances, only their *difference*. In addition, the inclusion of color legends in each trial may affect the predictive utility of color models. If our experiments were re-run without a visible color legend – such that subjects must make similarity judgments based on perception alone – it is possible that the results might align more closely with color model predictions. We leave exploration of these possibilities to future research.

DISCUSSION AND FUTURE WORK

In this work we evaluated nine quantitative colormaps using a relative similarity judgment task across varied spans of the data domain. We found that more recent multi-hue colormaps created using the CAM02-UCS color space – particularly *viridis* – perform well in terms of time and error. Single-hue colormaps perform well for larger data spans (i.e., judgments made over larger scale ranges), but exhibit issues of insufficient resolution at smaller spans. These results suggest that, by ramping in both luminance and hue, multi-hue colormaps can provide improved discrimination while preserving perception of order. We found that a diverging *blueorange* colormap performs similarly to the single-hue colormaps from which it is composed, but exhibits increased error for comparisons that straddle the mid-point. Finally, we confirmed that a rainbow colormap (*jet*) does indeed perform the worst overall in terms of both time and error, and should be jettisoned.

Our results provide actionable guidance for colormap design and selection. First, we establish benefits for judiciously designed multi-hue colormaps. In situations involving use of a continuous color scale to visualize a scalar field (e.g., in heatmaps), multi-hue colormaps may be preferable to single-hue given their improved resolution. For applications involving discrete color scales (i.e., with 5-7 colors), single-hue colormaps may still be acceptable; however, using a larger number of bins can result in color differences that fall within the low-span conditions studied here.

Second, we identify issues with low luminance regions set against a white background. Across colormaps (*greys*, *magma*, *plasma*), we observed much higher error rates despite similar distance estimates from perceptual color space models. We advise designers to avoid using these colormaps in situations with a high-luminance background, and warn that similar issues may arise when visualizing data using high-luminance colors against a dark background.

In a subsequent modeling exercise, we found that a combination of perceptually-uniform color models and categorical effects due to color naming can more accurately predict user performance than either alone. Larger perceptual and categorical differences correlate with improved accuracy, though with slightly diminishing effects for extreme differences. However, more work is needed to form more accurate models if we wish to advance automated colormap design and evaluation.

One limitation of the present work arises from our exchange of experimental control for ecological validity: through MTurk, we give up control of the viewing environments, the visual angle of the stimuli, along with other situational factors that confound color perception. Another limitation comes from our choice to present isolated color patches on a white background. Though white backgrounds are the most common both in print and on screen, our current setup is limited in its scope. Our experiments might be extended to other backgrounds, for example to see if analogous performance degradation occurs for light colors set in a dark context.

We chose to conduct an experiment on triplet comparison tasks in an abstracted context, configured to align with a standard observer model. However, visualizations in the wild involve a larger array of simultaneously presented colors, often involving variably sized marks across a variety of spatial configurations, and used for multiple perceptual tasks. These differences may very well affect colormap performance, for example due to simultaneous contrast. Similarly, while many of our findings likely still hold in scalar field visualizations, dedicated experiments in scalar field contexts might uncover additional effects of spatial frequency and gradients. Though our results provide actionable insights regarding the performance of colormaps in comparison tasks, future work might extend the findings to more real-world visualization examples.

ACKNOWLEDGMENTS

We thank both the anonymouse reviewers and members of the UW Interactive Data Lab for their helpful comments. This work was supported by a Paul G. Allen Family Foundation Allen Distinguished Investigator Award and a Moore Foundation Data-Driven Discovery Investigator Award.

REFERENCES

1. Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal.

Journal of Memory and Language 68, 3 (2013), 255 – 278.

2. Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48.
3. Lawrence D. Bergman, Bernice E. Rogowitz, and Lloyd A. Treinish. 1995. A rule-based tool for assisting colormap selection. In *Proc. IEEE Vis*. 118–125.
4. David Borland and Russell M. Taylor II. 2007. Rainbow color map (still) considered harmful. *IEEE Computer Graphics and Applications* 27, 2 (2007), 14–17.
5. Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3 data-driven documents. *IEEE Trans. on Visualization and Comp. Graphics* 17, 12 (2011), 2301–2309.
6. Cynthia A. Brewer. 1994. Color use guidelines for mapping. *Visualization in Modern Cartography* (1994), 123–148.
7. Cynthia A. Brewer. 1997. Evaluation of a model for predicting simultaneous contrast on color maps. *The Professional Geographer* 49, 3 (1997), 280–294.
8. Cynthia A. Brewer, Alan M. MacEachren, Linda W. Pickle, and Douglas Herrmann. 1997. Mapping mortality: Evaluating color schemes for choropleth maps. *Annals of the Association of American Geographers* 87, 3 (1997), 411–438.
9. Alžběta Brychtová and Arzu Çöltekin. 2017. The effect of spatial distance on the discriminability of colors in maps. *Cartography and Geographic Information Science* 44, 3 (2017), 229–245.
10. Robert C. Carter and Louis D. Silverstein. 2010. Size matters: Improved color-difference estimation for small visual targets. *Journal of the Society for Information Display* 18, 1 (2010), 17–28.
11. Michel E. Chevreul. 1860. *The Principles of Harmony and Contrast of Colours, and Their Applications to the Arts* (3 ed.). Henry G. Bohn, York Street, Covent Garden, London.
12. Jason Chuang, Maureen Stone, and Pat Hanrahan. 2008. A probabilistic model of the categorical association between colors. In *Color and Imaging Conference*, Vol. 2008. 6–11.
13. Johnson Chuang, Daniel Weiskopf, and Torsten Möller. 2009. Energy aware color sets. *Comp. Graphics Forum* 28, 2 (2009), 203–211.
14. William S. Cleveland and Robert McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. Amer. Statist. Assoc.* 79, 387 (1984), 531–554.
15. Gustav T. Fechner, Edwin G. Boring, Davis H. Howes, and Helmut E. Adler. 1966. *Elements of Psychophysics*. Holt, Rinehart and Winston.
16. Connor C. Gramazio. 2016. CIECAM02 Color. (2016). http://gramaz.io/d3-cam02/
17. Connor C. Gramazio, David H. Laidlaw, and Karen B. Schloss. 2017. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Trans. on Visualization and Comp. Graphics* 23, 1 (2017), 521–530.
18. Mark Harrower and Cynthia A. Brewer. 2003. ColorBrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal* 40, 1 (2003), 27–37.
19. Christopher G. Healey. 1996. Choosing effective colours for data visualization. In *Proc. IEEE Vis*. 263–270.
20. R. B. Heath, A. M.and Flavell. 1985. *Colour coding scales and computer graphics*. Springer Japan, Tokyo, 307–318.
21. Jeffrey Heer and Maureen Stone. 2012. Color naming models for color selection, image editing and palette design. In *Proc. ACM Human Factors in Computing Systems*. 1007–1016.
22. Gordon Kindlmann, Erik Reinhard, and Sarah Creem. 2002. Face-based luminance matching for perceptual colormap generation. In *Proc. IEEE Vis*. 299–306.
23. Helga Kolb, Ralph Nelson, Eduardo Fernandez, and Bryan Jones. 1995. *The Organization of the Retina and Visual System*. University of Utah Health Sciences Center, Salt Lake City, UT.
24. C. I. D. L'Eclairage. 1977. CIE recommendations on uniform color spaces, color-difference equations, and metric color terms. *Color Research & Application* 2, 1 (1977), 5–6.
25. Sungkil Lee, Mike Sips, and Hans-Peter Seidel. 2013. Perceptually driven visibility optimization for categorical data visualization. *IEEE Trans. on Visualization and Comp. Graphics* 19, 10 (2013), 1746–1757.
26. Adam Light and Patrick J. Bartlein. 2004. The end of the rainbow? Color schemes for improved data graphics. *Eos, Trans. American Geophysical Union* 85, 40 (2004), 385–391.
27. Sharon Lin, Julie Fortuna, Chinmay Kulkarni, Maureen Stone, and Jeffrey Heer. 2013. Selecting semantically-resonant colors for data visualization. In *Proc. IEEE EuroVis*. 401–410.
28. M. Ronnier Luo, Guihua Cui, and Changjun Li. 2006. Uniform colour spaces based on CIECAM02 colour appearance model. *Color Research & Application* 31, 4 (2006), 320–330.
29. M. Ronnier Luo, Guihua Cui, and B. Rigg. 2001. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research & Application* 26, 5 (2001), 340–350.

30. R. McDonald and K J Smith. 1995. CIE94-a new colour-difference formula. *Journal of the Society of Dyers and Colourists* 111, 12 (1995), 376–379.

31. Judy M. Olson and Cynthia A. Brewer. 1997. An evaluation of color selections to accommodate map users with color-vision impairments. *Annals of the Association of American Geographers* 87, 1 (1997), 103–134.

32. Stephen E. Palmer, Karen B. Schloss, and Jonathan Sammartino. 2013. Visual Aesthetics and Human Preference. *Annual Review of Psychology* 64, 1 (2013), 77–107.

33. Terry Regier and Paul Kay. 2009. Language, thought, and color: Whorf was half right. *Trends in Cognitive Sciences* 13, 10 (2009), 439 – 446.

34. Katharina Reinecke, David R. Flatla, and Christopher Brooks. 2016. Enabling designers to foresee which colors users cannot see. In *Proc. ACM Human Factors in Computing Systems*. 2693–2704.

35. Bernice E. Rogowitz and Alan D. Kalvin. 2001. The "Which Blair project": a quick visual method for evaluating perceptual color maps. In *Proc. IEEE Vis*. 183–556.

36. Bernice E. Rogowitz and Lloyd A. Treinish. 1998. Data visualization: the end of the rainbow. *IEEE Spectrum* 35, 12 (1998), 52–59.

37. Samuel Silva, Beatriz Sousa Santos, and Joaquim Madeira. 2011. Using color in visualization: A survey. *Computers & Graphics* 35, 2 (2011), 320 – 333.

38. Nathaniel Smith and Stéfan van der Walt. 2015. A Better Default Colormap for Matplotlib. (2015). https://www.youtube.com/watch?v=xAoljeRJ3lU

39. Maureen Stone, Danielle A. Szafir, and Vidya Setlur. 2014. An engineering model for color difference as a function of size. *Color and Imaging Conference* 2014, 6 (2014), 253–258.

40. Danielle A. Szafir. 2018. Modeling color difference for visualization design. *IEEE Trans. on Visualization and Comp. Graphics* 24, 1 (2018), 392–401.

41. Lujin Wangg, Joachim Giesen, Joachim Giesen, Peter Zolliker, and Klaus Mueller. 2008. Color design for illustrative visualization. *IEEE Trans. on Visualization and Comp. Graphics* 14, 6 (2008), 1739–1754.

42. Colin Ware. 1988. Color sequences for univariate maps: theory, experiments and principles. *IEEE Computer Graphics and Applications* 8, 5 (1988), 41–49.

43. Colin Ware. 2013. *Information Visualization: Perception for Design* (3 ed.). Morgan Kaufmann, Waltham, MA.

44. Colin Ware, Terece L. Turton, Francesca Samsel, Roxana Bujack, and David H. Rogers. 2017. Evaluating the perceptual uniformity of color sequences for feature Discrimination. In *EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3)*.

45. Jonathan Winawer, Nathan Witthoft, Michael C. Frank, Lisa Wu, Alex R. Wade, and Lera Boroditsky. 2007. Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences* 104, 19 (2007), 7780–7785.

46. Liang Zhou and Charles D. Hansen. 2016. A survey of colormaps in visualization. *IEEE Trans. on Visualization and Comp. Graphics* 22, 8 (2016), 2051–2069.