Colorgorical: Creating discriminable and preferable color palettes for information visualization

Connor C. Gramazio, *Student Member, IEEE*, David H. Laidlaw, *Fellow, IEEE*, Karen B. Schloss

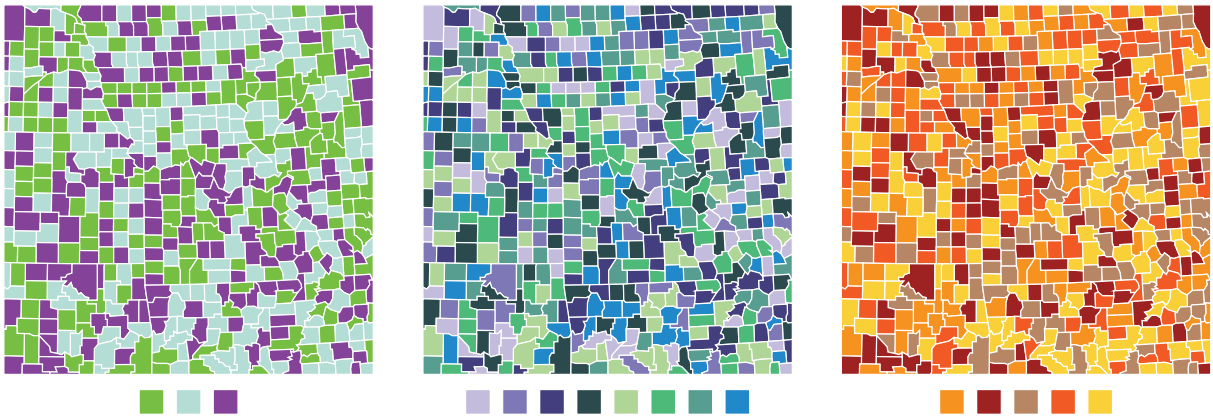


Fig. 1. Colorgorical palettes: (left) a 3-color palette generated with high name difference; (center) an 8-color palette generated with high pair preference; (right) a 5-color palette generated with high pair preference, medium perceptual distance, and a hue filter.

Abstract— We present an evaluation of Colorgorical, a web-based tool for creating discriminable and aesthetically preferable categorical color palettes. Colorgorical uses iterative semi-random sampling to pick colors from CIELAB space based on user-defined discriminability and preference importances. Colors are selected by assigning each a weighted sum score that applies the user-defined importances to Perceptual Distance, Name Difference, Name Uniqueness, and Pair Preference scoring functions, which compare a potential sample to already-picked palette colors. After, a color is added to the palette by randomly sampling from the highest scoring palettes. Users can also specify hue ranges or build off their own starting palettes. This procedure differs from previous approaches that do not allow customization (e.g., pre-made ColorBrewer palettes) or do not consider visualization design constraints (e.g., Adobe Color and ACE). In a Palette Score Evaluation, we verified that each scoring function measured different color information. Experiment 1 demonstrated that slider manipulation generates palettes that are consistent with the expected balance of discriminability and aesthetic preference for 3-, 5-, and 8-color palettes, and also shows that the number of colors may change the effectiveness of pair-based discriminability and preference scores. For instance, if the Pair Preference slider were upweighted, users would judge the palettes as more preferable on average. Experiment 2 compared Colorgorical palettes to benchmark palettes (ColorBrewer, Microsoft, Tableau, Random). Colorgorical palettes are as discriminable and are at least as preferable or more preferable than the alternative palette sets. In sum, Colorgorical allows users to make customized color palettes that are, on average, as effective as current industry standards by balancing the importance of discriminability and aesthetic preference.

Index Terms—Aesthetics in Visualization, Color Perception, Metrics & Benchmarks, Visual Design, Visualization

1 Introduction

We present an evaluation of Colorgorical (Fig. 2), a model and tool for creating arbitrarily sized, preferable, and discriminable color palettes for categorical information visualization (Fig. 1). As in other areas of design, it is important that a visualization color palette is aesthetically pleasing; but, unlike many other areas of design, visualization color palettes must also be highly discriminable. Balancing discriminability and aesthetic preference is challenging because they can be inversely related (i.e., preference increases with hue similarity [30], whereas discriminability decreases). Navigating this tradeoff requires design skill and experience, both beyond those of many visualization creators.

Colorgorical addresses this problem by operationalizing effective color palette selection with three color-scoring functions to balance discriminability and aesthetic preference: Perceptual Distance (CIEDE2000) [32], Name Difference [8], and a quantified model of color Pair Preference [30] (Sec. 3). A fourth, Name Uniqueness, was originally included, but was later removed because it had little effect on behavior (Sec. 6). With Colorgorical, color palette creation is simplified so that users need only specify the number of desired colors and drag sliders controlling color-scoring function importance to (1) create custom palettes that the average individual would find preferable while maintaining discriminability, and (2) explore how relative weights on discriminability vs. preference affect palette appearance. Users can further customize palettes by specifying desired hues and by building onto existing palettes (Sec. 4).

We evaluated Colorgorical's effectiveness in four ways: (1) runtime benchmarks (Sec. 4), (2) discriminability and preference score analysis (Sec. 5), (3) human-subject evaluation of different model settings (Sec. 6), and (4) human-subject evaluation of Colorgorical compared to industry standards (Sec. 7). We make the following contributions:

- We provide a technique to generate custom color palettes via

- *C.C. Gramazio and D.H. Laidlaw are with the Dept. of Computer Science at Brown University. E-mail: {connor,dhl}@cs.brown.edu.*
- *K.B. Schloss is with the Dept. of Cognitive, Linguistic, and Psychological Sciences at Brown University. E-mail: karenschloss@gmail.com.*

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

user-defined importance of discriminability and preference
- We detail the relations between Perceptual Distance, Name Difference, Name Uniqueness, and Pair Preference scoring functions
- We show how varying the relative weights of discriminability and preference sliders affects human discrimination performance and preference ratings
- We present evidence that Colorgorical palettes are as discriminable and often more preferable than industry standard, professionally hand-made color palettes

Colorgorical combines three features, making it a novel approach to palette design. First, it is designed specifically for visualization rather than for general art and design applications. Second, it uses empirically derived color preference data to inform categorical palette generation [30]. Third, it approaches visualization palette design by balancing categorical palette discriminability and preference.

2 Related Work

Current color palette tools are typically designed based on three types of strategies: discriminability optimization, color-term association mapping, or harmonic template application. We describe these approaches and discuss how Colorgorical targets limitations of past research.

2.1 Palette discriminability methods

A key issue in palette discriminability is whether a graphical mark can be quickly and accurately identified. Healey demonstrated that this problem can be addressed by using palettes whose colors are named with the 10 Munsell hues and that maximize perceptual distance between colors (CIEDE1976, Sec. 3) [7]. Maxwell also developed a discriminability-based technique to create categorical color palettes for multidimensional datasets based on classification dissimilarity of categories [16]. These approaches created discriminable palettes, but each has multiple limitations for design more broadly: (1) they do not address aesthetics, (2) Healey's technique is constrained to 10 or fewer color terms, and (3) they define perceptual distance using Euclidean distance (Healey) or maximum scaled difference (Maxwell) in CIELAB color space, which can be problematic due to perceptual uniformity limitations [13] (i.e., the same distance can have different perceptual consequences depending on the sampled region).

Colorgorical addresses these issues by (1) considering aesthetics in addition to discriminability (Sec. 3) [30], (2) ing 153 crowdsourced color terms compared to the 10 Munsell hues in Healey's method, and (3) using an updated perceptual distance function (CIEDE2000) that improves perceptual uniformity in the distance metric [32].

The color-name associations in Colorgorical are based on Heer and Stone's color-name statistics (Sec. 3) [8], which are derived from color-name association frequencies from the 153 most commonly-used names from the XKCD color-name crowdsourcing survey [19]. Name Difference measures the difference in color-name association frequency distributions between two colors. For example, green and red colors have large name differences because green colors have few associations with red names and vice versa. Presumably Name Difference is related to Perceptual Distance, but it is possible that they differ systematically, which we test in Sections 5 and 6. Name Salience, which we call Name Uniqueness to avoid confusion with color salience, captures the degree to which a color is specifically named (highly associated with only a few colors) vs. broadly named (moderately associated with many colors) (Fig. 1 in Supp. Mat.).

Another approach to designing discriminable palettes is for color experts to make pre-defined palettes (e.g., ColorBrewer [6]). Typically made through iterative design, experts construct these palettes by selecting colors that are discriminable under a variety of viewing conditions (e.g., after photocopying) and that support specialized tasks (e.g., ColorBrewer's "Accent" palettes emphasize certain colors). Although pre-made palettes are easy to use, they do not give visualization creators design flexibility or customizability. And although guidelines for hand-designing palettes exist [37], a visualization creator might not want to spend time or effort to learn about palette design. Colorgorical addresses this problem by allowing customization while building in constraints on aesthetics and discriminability; however, we leave support for specialized palettes (e.g., accent colors) for future research.

2.2 Color-term tools

Another way to create categorical palettes is through color-term associations. Crowdsourcing and linguistics-based approaches can produce color-term associations that create semantically meaningful palettes (e.g., a "mango ice cream" category might produce a light orange) [11, 31]. Setlur and Stone show that various natural language processing techniques can be used to mine color-semantic pairings from large text datasets [31]. Colorgorical does not currently support semantic mappings, but it is an exciting future direction.

2.3 Harmonic template tools

Many harmony-based categorical color palette tools are targeted for general-purpose design and do not focus on visualization design constraints (e.g., discriminability). These tools create palettes based on harmony principles in color theory [20, 22]. A common implementation of harmony is through *harmonic templates* based on hue relations [15], such as the two-color complementary relation that stems from Itten's version of harmony (e.g., blue and orange) [10]. For example, Adobe Color creates 5-color palettes based on harmonic templates and optional image color analysis [21]. Similarly, Dial-a-color, uses harmonic templates as a starting point and allows users to alter color properties like lightness and saturation [18]. ACE lets users manipulate discrimination and harmony importance for interface design by answering a series of questions in a text interface about each colored interface component [17] (unlike ACE, Colorgorical is not limited to interface coloration and uses sliders to balance discrimination and aesthetic preference rather than a text interface). Finally, the Harmonious Color Scheme Generator constructs color palettes through *familial factors* (promoting similarity along hue, saturation, or lightness dimensions) and *rhythmic spans* (sampling colors using a fixed uniform interval along a color dimension) [9].

Harmonic templates were generated from color theory in art without empirical validation [10], and do not necessarily correspond to human judgments of harmony. For example, the notion that complementary colors are harmonious is key to the notion of harmonic templates. Yet humans judge complementary hues as among the least harmonious and instead judge more similar hues as more harmonious [23, 24, 30, 34].

Although the term "harmony" is often used interchangeably with aesthetic preference [2], the two are not the same [24, 30]. Schloss and Palmer demonstrated how they differ, where harmony was defined as "how strongly an observer experiences the colors in the combination as going or belonging together, regardless of whether the observer likes the combination or not," and preference is "how much an observer likes a given pair of colors as a Gestalt, or whole" [30]. Although both increased with hue similarity, pair preference relied more on preference ratings for individual colors and on lightness contrast, whereas harmony relied more on desaturation (i.e., pairs with less saturated colors were more harmonious).

Colorgorical uses Schloss and Palmer's pair preference model (Sec. 3) [30] rather than harmony because we reasoned that how much people like visualization palette colors is more central to the present aims than how well they feel the colors go together.

3 Background: Model Scoring Functions

Colorgorical iteratively samples colors using three color discriminability scores (*Perceptual Distance*, *Name Difference*, *Name Uniqueness*) and a color preference score (*Pair Preference*). Colorgorical assumes that discriminability and preference for large combinations of colors can be predicted by these lower-order scores. Name Uniqueness was ultimately removed from the model because it had little effect on discriminability performance or preference (Sec. 6).

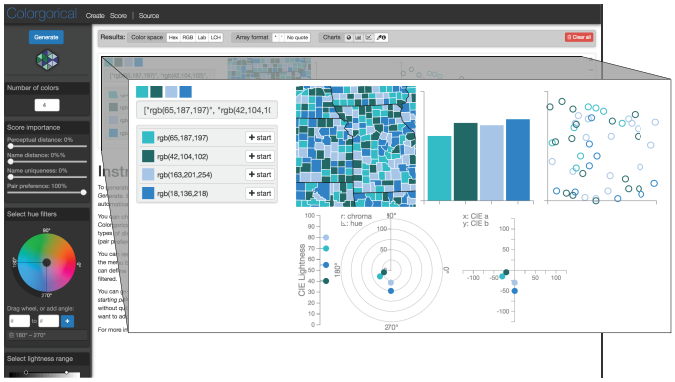


Fig. 2. A Colorgorical screenshot. Here a user has specified a hue filter (left) and has generated a 4-color palette (detail). Users can list colors in many color spaces and render colors in a variety of charts.

Each score operates in CIELAB. The $L*$ axis of CIELAB approximates a color's lightness, the $a*$ axis approximates its redness-to-greenness, and the $b*$ axis approximates its blueness-to-yellowness. To support Name Difference and Name Uniqueness, we use a modification of CIELAB that quantizes the space into 8,325 discrete colors by sampling every 5 units along each axis starting at the origin [8]. Some scores also depend on CIE LCh, which is a polar representation of the Euclidean CIELAB space. In CIE LCh, $L*$ is the same as in CIELAB, but the $a*$ and $b*$ axis are converted to chroma (C, radius) and hue (h, angle).

3.1 Color discriminability scores

We used multiple discriminability scores because perceptual difference might differ from name difference. For instance, a chartreuse (yellow-green) might be perceptually distinct from a green or yellow but might be called green or yellow, making it easy to confuse with other greens or yellows in a visualization when referenced by name.

3.1.1 CIEDE2000: Perceptual Distance

To calculate Perceptual Distance between two colors we use CIEDE2000 (DE_{00}) [32]. It is similar to the original CIEDE, DE_{76} (Euclidean CIELAB), but DE_{00} calculates distance in CIE LCh with a hue rotation term (R_T) and corrections for lightness (S_L), chroma (S_C), and hue (S_h) to improve perceptually uniformity [14].

$$\mathrm{DE}_{76} = \sqrt{\Delta L^2 + \Delta a^2 + \Delta b^2} \tag{1}$$

$$\mathrm{DE}_{00} = \sqrt{\left(\frac{\Delta L}{S_L}\right)^2 + \left(\frac{\Delta C}{S_C}\right)^2 + \left(\frac{\Delta H}{S_H}\right)^2 + R_T \frac{\Delta C}{S_C}\frac{\Delta H}{S_H}} \tag{2}$$

3.1.2 Name Difference

Name Difference (ND) captures the degree to which two colors have distinct color-name association frequency distributions [8]. Color-name associations are mappings between colors and names (e.g., `rgb(255,0,0)` → "bright red"). The name data are composed of the discretized CIELAB color space (C) described earlier, a list of 153 popular color names (W), and a color-name association frequency matrix (T) that has C rows and W columns. The scores also rely on the conditional probability of a color name w given any color in C:

$$p(w|c) = T_{c,w} / \sum_{w} T_{c,w} \tag{3}$$

We calculate Name Difference using Hellinger distance [8]:

$$\mathrm{ND}(c_1, c_2) = \sqrt{1 - \sum_{w \in W} \sqrt{p(w|c_1)p(w|c_2)}} \tag{4}$$

3.1.3 Name Uniqueness

Name Uniqueness (NU) captures the degree to which colors have uniform distributions of color-name association frequencies. Colors that have few strongly associated names (i.e., a focal distribution) result in lower scores, whereas colors that have many weakly associated names (i.e., a more-uniform distribution) result in higher scores. Name Uniqueness is calculated by using the negative entropy of a color's name-association frequency distribution from the color-name-association matrix (T) and the list of color names (W):

$$\mathrm{NU}(c) = -\mathrm{H}(p(W|c)) = \sum_{w \in W} p(w|c) \log p(w|c) \tag{5}$$

Unlike the other two discriminability measures, Name Uniqueness relies on individual colors rather than relations between other colors within the palette. We believe this is a key reason why it was not useful in the Colorgorcial model (Sec. 6).

3.2 Aesthetic preference score: Pair Preference

Pair Preference (PP) is based on a linear regression model used to predict pair preferences from three color-appearance and color-relation factors, which was previously operationalized in Munsell space [30]. The best-fit model explained 53.5% of the variance in pair preference judgments with three factors: coolness (κ), hue similarity ΔH, and lightness contrast ΔL. We have altered the original equation to use CIE LCh rather than Munsell color space coordinates, as is reflected in the hue similarity and lightness contrast terms[1]. Coolness scores are calculated in CIE LCh using a linear interpolation of the original 32 color-coolness mappings, which approximates the number of hue-steps a color is from Munsell 10R, such that greenish blues are cool and orangish reds are not cool (Supp. Mat.). The Pair Preference scoring function reflects people's preference for color combinations that contain cool colors that differ in lightness and are similar in hue.

$$\mathrm{PP}(c_1, c_2) = 75.15(\kappa_1 + \kappa_2) + 47.61|\Delta L| - 46.42|\Delta H| \tag{6}$$

4 Colorgorical Model

Colorgorical generates color palettes using iterative semi-random sampling. Users specify the number of desired colors and use sliders to set the relative balance of aesthetic preference and discriminability (Sec. 3). Generated palettes are displayed to the user as a swatch, map, bar chart, and scatterplot, which highlights how the discriminability may shift with different types and sizes of graphical marks [4, 33].

4.1 Minimum discriminability & preference assertions

Each palette is built from an 8,325-color discretized D65 CIELAB space (Sec. 3) and is additionally filtered in three ways to help increase discriminability and preference, which we describe below: (1) noticeable difference; (2) lightness clamping (from $L* = 25$ to $L* = 85$) and (3) filtering the dark yellow (generally disliked) region of color space. Although the same RGB coordinates can result in different CIELAB colors on different monitors if monitors are uncalibrated, Stone et al. show that using a fixed correspondence between D65 CIELAB and RGB can be used effectively for online tools in practice [33].

Discriminability The model enforces a lower discriminability bound by sampling *noticeably different colors* using Stone et al.'s noticeable difference function, which provides a minimum CIELAB interval required to discriminate the colors of two graphical marks more than 50% of the time (based on their physical size) [33]. We use a small, conservative visual angle in our calculations ($1/3^\circ$) and multiply the function's suggested interval by three for extra caution.

To ensure discriminability we also exclude colors that are lighter than $L = 85$ and darker than $L = 25$ so that all colors are visible on black or white backgrounds ($L_{\text{black}} = 0$, $L_{\text{white}} = 100$). Colorgorical only includes RGB-valid colors.

[1]The CIE LCh model explains 51.8% of the variance in Schloss and Palmer's preference data (their Munsell-based model explains 53.5%).

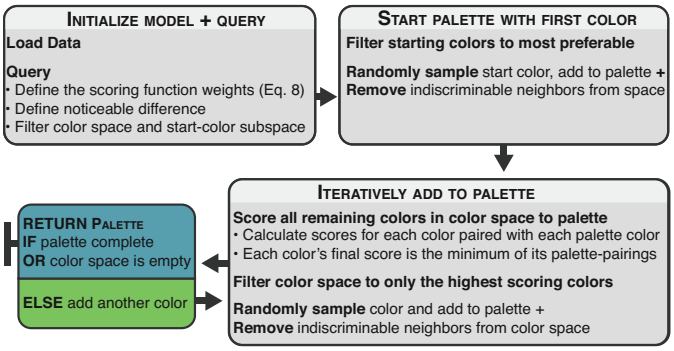


Fig. 3. Diagram of Colorgorical palette construction procedure.

Preference Colorgorical excludes the dark yellowish-green region of CIE LCh, which has strongly disliked colors, on average, across many cultures [26, 35, 36]. We define this region as $L \in [35, 75]$ and $H \in [85°, 114°]$. While there are individual differences in preference [25, 30] and some observers may like these colors [29], the goal is to cater to the average observer. This filter was especially important for generating aesthetically preferable discriminable palettes because of the way Pair Preference and discriminability functions interact. In the Pair Preference equation, the coolness term biases selection toward bluish hues and the lightness term biases selection of contrasting lightness. The discriminability functions bias selection for colors that are far apart in CIELAB color space (i.e., contrasting hue and lightness). Once bluish hues are selected, discriminability would be promoted in subsequent color selections by selecting opposite, yellowish hues of a different lightness level (opposite ends of the $b*$ and $L*$ axes). If the blues are remotely light, then selected yellows will be the dark yellows that people generally dislike. The removal of this region still retained a large region of color space that was sufficiently discriminable to pair with blues, while increasing typical aesthetic palette preference.

To maximize preference within a defined balance, the model generates 10 palettes and returns the palette with the highest minimum-Pair-Preference given all color pairings in each palette.

4.2 User-defined model parameters

In addition to specifying the number of colors and manipulating discriminability and preference sliders, users can also configure two optional parameters. First, they can limit color sampling to certain hue ranges (e.g., reds only, or reds and blues), which supports tasks such as designing around brand colors. Second, users can supply an existing palette for Colorgorical to build on. If users provide a palette, Colorgorical rounds the input to the nearest quantized CIELAB color and adds new colors until the palette reaches the desired size.

4.3 Palette construction process

Palettes are generated in three steps: (1) initialize, (2) start a palette with the first color, and (3) iteratively add new colors (Fig. 3). Colorgorical can typically generate palettes with up to 22 colors before exhausting color space. However, it is inadvisable to use that many colors due to perceptual limitations [5]. If no more colors can be sampled, Colorgorical returns a partial palette and an error message.

4.3.1 Step 1: Initialize

Initialization starts by loading CIELAB space, color coolness scores, and color-name associations into memory. A CIELAB subspace is also loaded into memory, which samples every 15 units along each CIELAB axis and is used along with a precomputed Pair Preference score matrix to pick the first palette color. We use a coarser subspace to select the first color because using precomputed Pair-Preference scores for all pairs of 8,325-colors takes too long for interactivity due to combinatorial explosion. Color space can be filtered based on parameters provided by the user (e.g., hue filters). After applying optional filters, the model limits the subsampled space colors (c) and the color pair preference matrix (Φ) to highly preferable colors (i.e., no dark yellows) using a standard deviation (SD) preference threshold (Eq. 7). The threshold removes any color-pair row from Φ whose pair preference score is less than the standard deviation-based limit. Then, the starting color is sampled from the unique colors remaining in Φ's color-pair rows.

$$\text{threshold}(c) = \Phi_c > \max(\Phi) - 0.75 * \text{SD}(\Phi) \tag{7}$$

The last initialization step also defines a noticeable difference with Stone et al.'s CIELAB intervals described above, which removes colors that are too similar to each sampled color. Sampled color differences must have at least one axis above the following intervals: $\Delta_L = 22.747, \Delta_a = 31.427, \Delta_b = 44.757$.

4.3.2 Step 2: Start palette

The first color of a palette is selected by randomly sampling a seed color from the remaining colors after Step 1. Next, all colors that are not noticeably different from the seed are removed from color space using the CIELAB intervals defined in Step 1. Sampling is skipped if users provide their own seed color(s), but indiscriminable neighboring colors are still eliminated.

4.3.3 Step 3: Add to palette

To add a new color, the model computes scores for all remaining colors using a weighted sum (Eq. 8). This function sums each of the four minimum palette scores ($\vec{\Psi}$) with user-defined weights ($\vec{w}$), given all possible scores between a potential new color (c) and the already picked colors (P). The model uses minimum palette scores assuming that a palette is only as discriminable or preferable as its lowest score. There is also a hue-dependent penalty term (τ) to reduce the likelihood of sampling a color bordering the dark yellow filter region. The new color is then randomly sampled from colors that fall above a score threshold (Eq. 7, where Φ is now weighted-sum scores). Non-discriminable colors are removed after sampling.

$$\text{score}(c, P) = \tau(\vec{w} \cdot \vec{\Psi})$$
$$\tau = \begin{cases} 0.75, & \text{if } 115° < c_{\text{hue}} < 138° \wedge c_{\text{L}} \leq 45 \\ 0.8, & \text{if } 70° \leq c_{\text{hue}} \leq 115° \wedge 45 < c_{\text{L}} \leq 75 \\ 0.85, & \text{if } 70° \leq c_{\text{hue}} \leq 115° \wedge c_{\text{L}} > 75 \\ 1, & \text{otherwise} \end{cases} \tag{8}$$

4.4 Implementation and Performance

Colorgorical is implemented in C-accelerated Python. To evaluate average model runtime (50 runs) as a function of palette size (1 to 20 colors), we profiled single-palette generation on a Mid 2012 MacBook Pro Retina with a 2.6 GHz Intel Core i7 CPU and 16GB 1600MHz DDR3 RAM. Average initialization time was 140ms (SEM = 0.004). If a palette was returned before reaching the required number of colors, it was discarded and the test was run again. Runtime performance increased linearly in the number of colors such that adding a color increased runtime by 17.6ms on average (Supp. Mat.).

5 Palette Score Evaluation

Before conducting human-subject testing, we first tested whether any of Colorgorical's scoring functions (i.e., Perceptual Distance, Name Difference, Name Uniqueness, and Pair Preference) could be removed from the model to simplify its design without significantly affecting palette output. For instance, if Perceptual Distance were to explain most of the variance in Name Difference scores, then the Name Difference scoring function could be removed from the model with little effect on palette output.

We examined the similarity among the four Colorgorical scoring functions using multiple linear regressions to predict 39,600 *palette scores* for each *palette scoring function* (Sec. 3) from the three remaining functions (e.g., predicting Perceptual Distance from Name Difference, Name Uniqueness, and Pair Preference). *Palette scores* are the minimum palette scoring function output given all color pairs in a

palette. We use the minimum score because we assumed that a palette is only as preferable or discriminable as its lowest pair. The number 39,600 stems from the full range of possible Colorgorical slider settings and 3 palette sizes (66 settings, {3,5,8}-colors, 200 repeats). The 66 settings were made from the different unique combinations from dragging each of the four sliders to 0%, 50%, or 100%, which ignore duplicate settings encountered when moving one or more sliders to 0%.

We also examined how the four palette scores changed with palette size. Below we highlight results and implications from our analyses, and the methods and full results are in Supplementary Material.

Both Perceptual Distance and Name Difference were strong positive predictors of one another. Name Uniqueness was always a weak negative predictor of the other scores. Pair Preference was always a strong negative predictor of Perceptual Distance and Name Difference. Further, Pair Preference was more strongly related to Name Difference than to Perceptual Distance. Given that palette scores in each palette scoring function were significantly predicted by all three of the other palette scoring functions, we concluded that each scoring function measured sufficiently different color information to justify keeping them all in the model for Experiment 1.

6 Exp. 1: Model human-subject evaluation

Experiment 1 tested how palette discriminability performance and preference ratings varied as the relative weights on the Colorgorical sliders varied (i.e., the relative importance of each scoring function; Sec. 4). We also identified which slider settings produced the most discriminable or preferable palettes to prepare for a comparison between Colorgorical and current industry standards in Experiment 2 (Sec. 7).

Experiment 1 used the same representative palettes as in Section 5, which were analogous to the slider settings produced by moving each to either 0%, 50%, or 100% for 3-, 5-, and 8-color palettes.

Discrimination performance and preference were assessed using two difference tasks (Fig. 4). In the discrimination task, participants reported which side of a map had more counties of a target color, providing data on number of errors and response time (RT). In the aesthetic preference task, participants rated how much they liked the color combinations in each palette. We predicted that:

P1 Palettes with fewer color would be more discriminable
P2 Discrimination RT and error would correlate in a strong negative direction with Perceptual Distance and in a strong positive direction with Pair Preference, whereas preference ratings would show the opposite pattern
P3 Palette size would modulate the discriminability and preference ratings associated with each slider setting.
P4 Slider settings would significantly predict discrimination performance and preference ratings

P1 is based on previous evidence that visualizations with more colors are harder to process [5]. **P2** extends Palette Score Evaluation findings that Perceptual Distance and Name Difference negatively predicted Pair Preference. **P3** builds on the first two predictions: based on **P1** we expect that palette size will modulate the discriminability of slider settings, and based on **P2** we expect that preference will be negatively correlated with discriminability. **P4** makes two strings of assumptions based on the Palette Score Evaluation: (1) the trade-off between discrimination and preference palette scores will extend to behavior (**P2**) and (2) the relative importance of scoring functions (i.e., slider settings) would affect behavior in the same manner as palette scores (e.g., a higher relative importance of Pair Preference will produce higher Pair Preference palette scores). By transitivity, we predict that slider settings will be indicative of behavior.

6.1 Methods

6.1.1 Participants

77 participants completed the discrimination task and 60 completed the preference rating task (recruited through Amazon Mechanical Turk, $3 compensation). Palette size (3-, 5-, 8-colors) was a between-subjects factor. For quality control, we determined *a priori* to discard participants who were $< 60\%$ accurate across all trials in the discriminability task (3-color: $n = 3$; 5-color: $n = 6$; 8-color: $n = 8$). No participants were discarded in the preference task. In the final datasets there were 20 participants per palette size in each task, and discard frequency did not significantly differ between palette size conditions ($\chi^2(2) = 1.793, p = 0.408$). All self-reported having normal color vision and gave informed consent. The Brown University IRB approved the experiment protocol.

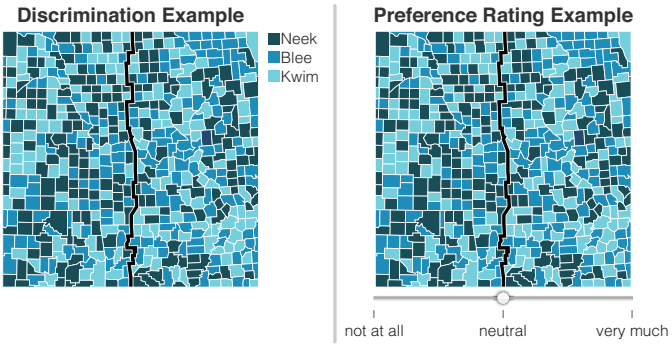


Fig. 4. Discrimination and preference rating task stimuli. The discrimination task asked users which side had more "Neek" counties (← and → keys). The preference rating task asked users to click on the slider.

6.1.2 Design & Displays

The experimental designs for the discrimination and preference tasks were similar. In both, each participant saw 660 palettes from 66 slider settings (see Sec. 5 for setting information) with 10 different color palettes within each slider setting (treated as repetitions). The specific colors in each palette varied across participants (simulating different runs of Colorgorical), but were generated with the same experimental design. Palette size varied between-subjects (3, 5, or 8 colors).

The palettes that comprised the displays for the discrimination task were also used for the preference task, such that each discrimination participant was yoked to a preference participant (i.e., both saw the same palettes). Palettes were displayed on a predefined map of 554 counties in the U.S. (300×300 pixels). The map itself differed slightly based on the task (Fig. 4).

For the discrimination task, a 5-pixel-wide contour bisected the map (adhering to county borders). The contour was black and the county borders were white so that both would fall outside of Colorlogical's default lightness sampling range ($L \in [25, 85]$; $L_{Black} = 0; L_{White} = 100$). The size of the counties on each side were slightly altered so they were approximately equal (left: 165 px; right: 163 px). A legend rendered to the right of each map assigned each palette color to a nonsense word category. The target "Neek" color was always at the top of the legend to prevent participants from having to search for the target color. One side of the map had over-represented target color ("Neek"; $1.5\times$ more frequent on one side than the base rate) and the opposite side had an over-represented distractor color ($1.3\times$ more frequent). The target side was left/right balanced across trials. Based on our assumption that a palette is only as effective as its least discriminable pair of colors, the target and distractor colors were always the palette colors with the lowest and second-lowest Perceptual Distance scores compared with all other colors in the palette, respectively.

In the preference task, there was no dividing contour and no legend, the colors were roughly equal in proportion, and they were randomly assigned to positions across the map (no left/right asymmetry). Below the map there was a 300-pixel-wide continuous response slider scale ranging from -100 to 100 with labeled extrema and midpoint (left: "not at all"; right: "very much"; midpoint: "neutral") [30]. The scale was initialized with the slider set to "neutral" to avoid biasing participants.

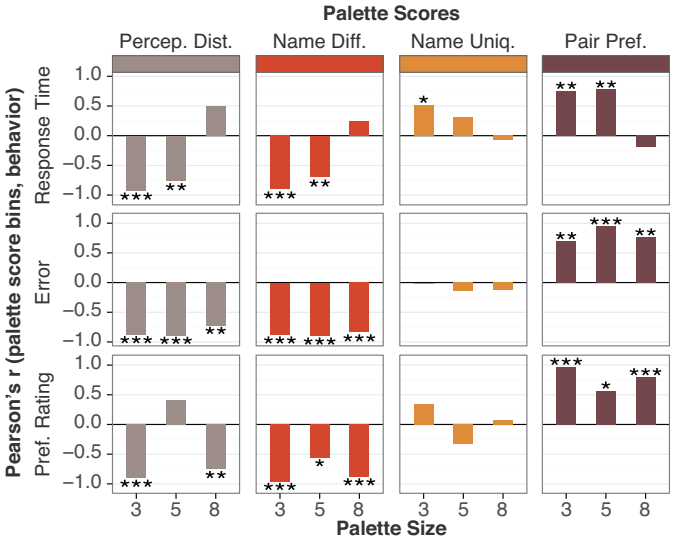


Fig. 5. Correlations between binned palette scores and responses on each measure for each palette size ($*: p < 0.05; **: p < 0.01; ***: p < 0.001$).

6.1.3 Procedure

Discrimination Task. Participants were first presented with an example display and were told that their task would be to indicate which half (left/right) of a map had more Neek counties using the left/right arrow keys. They were also told that the target Neek color would always be shown at the top of a legend and that answers would be marked incorrect if they did not respond within 3.5 seconds. Participants completed five practice trials using distinct displays from the 660 test maps they would see in the experiment, followed by 660 test trials. Maps were shown in random order in the center of the window. Trials were separated by a 500-ms inter-trial interval with a fixation cross displayed at the center of the screen. Optional breaks were given every 20 trials. This task took ~30 minutes to complete.

Preference Task. Participants were asked to rate their aesthetic preference for the color combination in each palette by clicking a point on a slider between the left ("not at all" preferable) and the right ("very much" preferable) ends (Fig. 4). To help them gauge what liking "not at all" and "very much" meant to them in the context of these color combinations, participants were shown an anchoring page containing 66 representative maps. They scrolled through the maps and considered how they would rate each map while using the full range of the scale. During the experiment, each map was presented one at a time in a random order (separated by a 250-ms blank pause screen). The preference slider appeared 1 second after the map appeared to encourage participants to consider their preference carefully before responding. This task took ~40 minutes to complete.

6.2 Results and Discussion

Before analysis, we pruned response time (RT) data by removing incorrect trials and then eliminating trials for each subject that were more than ± 2.5 standard deviations away from their mean RT [27]. On average, 129 errors (19.5%) and 24 outliers (3.6%) were removed.

Overall participant accuracy decreased as palette size increased (3-color average error: 79/660; 5-color: 119/660; 8-color: 190/660), indicating that displays with smaller color palettes were more discriminable (**P1**, **P3**). This result mirrored the increased participant discard rate for larger palette size conditions due to high error rates (Sec. 6.1.1) and is consistent with previous findings that showed visualizations with fewer color categories are more effective [5]. Between-subjects one-way ANOVAs testing for effects of palette size (3, 5, 8) within each measure indicated significant effects for number of errors ($F(2,57) = 30.801, p < .001$) but not for RT or preference ($F(2,57) = 1.574, 1.035; p = 0.216, 0.362$, respectively).

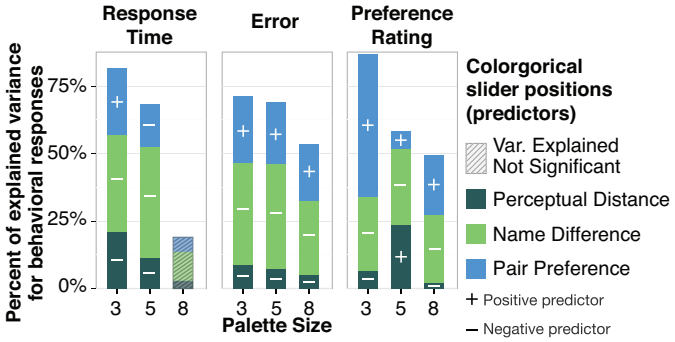


Fig. 6. Variance explained (R^2) for the 9 models decomposed to look at the variance explained of behavioral data in terms of slider settings (i.e., palette score relative importance).

6.2.1 Palette score and behavioral measure correlations

Figure 5 shows the correlations between each type of palette score (Perceptual Distance, Name Difference, Name Uniqueness, and Pair Preference) and the three behavioral measures (RT, error rate, and preference ratings), averaged over participants. *Palette score* refers to the lowest *palette scoring function* value (Sec. 3) given all color pairs in a palette. To conduct these analyses, we first binned the behavioral data for each measure according to palette scores (15 equally-spaced bins) for each subject[2]. After, we averaged the data for all palettes that scored in the same bin and then averaged those values across participants. This binning was necessary prior to averaging across participants because each participant saw different palettes with slightly different scores (Supp. Mat.). For example, RT for palettes with a Pair Preference scores of 30.03 and 30.05 would be binned together.

We cross-checked the binned-score correlations by calculating the within-subject correlations for each behavioral measure and palette score and then used Fisher's Z transform prior to calculating the between-subject average Pearson's r for each measure and score combination. For the most part, these analyses showed the same pattern of results as the binned correlation statistics (Supp. Mat.).

The binned-score correlations are presented below (see Supp. Mat. for individual correlations on non-binned data). In summary, the Perceptual Distance, Name Difference, and Pair Preference scores had the predicted effects: RT and error rates decreased (i.e., better performance) as Perceptual Distance and Name Difference increased, but they increased (i.e., worse performance) as Pair Preference increased (**P2**). In contrast, preference decreased as Perceptual Distance and Name Difference increased and they increased as Pair Preference increased. Name Uniqueness had little to no effect.

RT. RT decreased as Perceptual Distance and Name Difference increased for 3- and 5-color palettes (Perceptual Distance: $r(13) = -0.926, r(13) = -0.757; p \leqslant 0.001$ respectively; Name Difference: $r(13) = -0.893, r(13) = -0.689; p \leqslant 0.005$ respectively). Similarly, Pair Preference followed **P2** for 3- and 5-color palettes with strong positive correlations with RT ($r(13) = 0.755, 0.776; p = 0.001$). Name Uniqueness was significantly correlated with RT for 3-colors ($r(13) = 0.521; p = 0.046$), but not for 5-colors ($r(13) = 0.306; p = 0.268$). No scores were significantly correlated with RT for 8-color conditions ($r(12) = 0.496, r(13) = 0.251, -0.071, -0.193; p \geqslant 0.071$ for Perceptual Distance, Name Difference, Name Uniqueness, and Pair Preference respectively). These findings largely support **P2** for 3 and 5 colors; however, 8-color palette correlations were not significant.

Error. Error rate correlations were significant for all sizes with Perceptual Distance ($r(13) = -0.887, -0.898, r(12) = -0.731; p \leqslant 0.003$, for 3-, 5-, and 8-colors respectively), Name Difference ($r(13) = -0.874, -0.892, -0.838; p < 0.001$), and Pair Preference ($r(13) = 0.697, 0.945, 0.761; p \leqslant 0.004$). Similar to RT correlations, Name

[2]The degrees of freedom for 8-color perceptual distance correlations is one less due to an empty bin, which is shown in the Supplementary Material.

Uniqueness was not significantly related to error measures ($r(13) = -0.016, -0.141, -0.126; p \geqslant 0.616$).

Preference Rating. Preference rating trends were the opposite of error and RT, and consistent with **P2**. Increasing 3- and 8-color Perceptual Distance reduced preference ratings, ($r(13) = -0.897, r(12) = -0.751; p \leqslant 0.002$) but not significantly so for 5-color palette ($r(13) = 0.412; p = 0.127$). Preference ratings also decreased as Name Difference increased for 3-, 5-, and 8-colors ($r(13) = -0.969, -0.57, -0.891; p \leqslant 0.026$, respectively). Increasing Pair Preference increased preference ratings ($r(13) = 0.971, 0.57, 0.796; p \leqslant 0.026$). Again, Name Uniqueness was not significantly related ($r(13) = 0.346, -0.333, 0.073; p \geqslant 0.207$).

6.2.2 Predicting behavioral measures from slider settings

To test whether slider settings (i.e., relative importance of the *palette scoring functions*) significantly predict behavior (**P4**), we performed a series of multiple linear regressions that predicted behavioral measures as a function of changing sliders to 0%, 50%, or 100% (Fig. 6). Given that the correlational analyses above suggested that Name Uniqueness had little effect on behavior, we averaged slider configurations that would be equivalent if Name Uniqueness were ignored. For example, if Perceptual Distance and Name Uniqueness were both set to 50%, the new setting would be Perceptual Distance as 100% and would be averaged with other palettes where Perceptual Distance is 100%. This reduced the regression analysis to predict 20 unique slider settings rather than the previous 66. The data that were input to the correlations are graphed in the Supplementary Material.

Below we detail the results of the multiple linear regressions using slider relative importance to predict the three behavioral measures. More information about the relation between sliders, size, and behavioral measures is provided in the Supplementary Material. In summary, the slider settings were typically able to significantly predict the behavioral measures (**P4**).

RT. RT decreased (improved) as Perceptual Distance and Name Difference slider weights increased and RT increased (got worse) as Pair Preference slider weights increased (**P4**; Supp. Mat.). Name Difference was always the most predictive and Perceptual Difference and Pair Preference were similarly less predictive (Fig. 6). Although this pattern was present for all three palette sizes, the models were significant for the 3- and 5-color palettes ($F(3,16) = 23.442, 11.447; R^2 = 0.815, 0.682; p < 0.001$, respectively), but not the 8-color palettes ($F(3,16) = 1.267, R^2 = 0.192, p = 0.319$). The lack of significance for 8-color palettes coincides with the oddity that response time was typically faster for 8-color palettes than 5-color ones; this is unexpected, given (1) past visual search research finding that more colors take longer to discriminate [5] and (2) the previously discussed palette size relation with accuracy. We suspect that this difference may be because participants tried less hard or the task became too difficult in the 8-color condition because they had higher overall error rates. Another possibility is that pair-based color discriminability scores (e.g., Perceptual Distance) may break down as the number of colors increases, which would create a need for higher-order combination discriminability scores. Each of these possibilities raise interesting future directions for studying the relation between palette effectiveness and number of colors.

Error. Slider relative importance analysis mirrored RT (**P4**; Supp. Mat.), except that Pair Preference was more important than Perceptual Distance (Fig. 6 and Supp. Mat.). The reason for this difference is unknown. The multiple linear regressions for all 3-, 5-, and 8-colors were all significant ($F(3,16) = 13.186, 11.964, 6.192; R^2 = 0.712, 0.692, 0.537; p \leqslant 0.005$, respectively).

Preference Rating. Preference ratings increased with weights on the Pair Preference slider and decreased with weights on the Perceptual Distance and Name Difference sliders (**P4**; Supp. Mat. slider-behavior figure). Pair Preference was the most predictive slider for 3-colors, but not for 5- and 8-colors (Fig. 6 and Supp. Mat.); instead, Name Difference was most predictive. Perceptual Distance was more important than Pair Preference for 5-colors, but was otherwise the least important slider. The multiple linear regressions for 3-, 5-, and 8-colors were all significant ($F(3,16) = 35.089, 7.396, 5.228; R^2 = 0.868, 0.581, 0.495; p \leqslant 0.01$, respectively).

It is noteworthy that the model's ability to predict preference ratings decreased for 5-colors relative to 3-colors, suggesting that the mechanism behind human aesthetic preference ratings may deviate from pair-based preference predictions as the number of palette colors changes. Another difference for 5-color palettes, compared to 3- and 8-colors, was that all settings were rated either neutral or slightly negative. These results suggest that the assumption that pair-wise based preference models generalize to palettes of three colors might break down for larger palettes. The differences in preference ratings over palette sizes motivates the need for further research on the aesthetics of higher-order color combinations.

We also found that preference ratings decreased faster as Name Difference relative importance was increased compared to increases in Perceptual Distance relative importance (see Supp. Mat.). This asymmetry might be caused by differences in how Perceptual Distance and Name Difference measure distances in color space. It could be that Perceptual Distance is more supportive because it can generate color pairs that differ primarily in lightness (which is one of the terms in Pair Preference), whereas Name Difference might be more likely to favor differences in hue, which would be in opposition to Pair Preference's hue similarity term.

6.2.3 Lowest-Error and Highest-Preference settings

A main goal of Experiment 1 was to determine which Colorgorical settings to use to generate color palettes for comparison against current standards (Experiment 2). The combinatorial explosion of conditions prevented comparing all slider combinations to current standards. Therefore, we chose to select slider settings that either produced highly discriminable or highly preferable palettes (i.e., at either end of the previously-discussed discriminability-preference trade-off). Figure 8 shows the lowest discrimination error setting (subsequently called "Low-Error" palettes) and the highest preference rating setting ("Preferable" palettes) for each palette size. There were significantly fewer errors for Low-Error palettes than for Preferable palettes ($t(19) = 3.322, 7.589, 3.15; p \leqslant 0.005$, 3-,5-,8-colors). Preference ratings were significantly greater for Preferable palettes than for Low-Error palettes for 3- and 8-colors ($t(19) = 4.610, 2.841, p \leqslant 0.01$), but not for 5-colors ($t(19) = 0.499, p = 0.623$) (consistent issues about 5-color palettes discussed above).

6.2.4 Summary

Experiment 1's results largely support each of our four predictions and suggest that Colorgorical's sliders are effective at controlling the discriminability and preference of color palettes, although some 5- and 8-color conditions led to unexpected behavioral results. Discriminability performance typically improved (faster RT, fewer errors) as the Perceptual Distance and Name Difference palette scores increased (and with greater weight on their corresponding sliders) and Preference judgments typically increased as Pair Preference palette scores increased (with greater weight on its slider) (**P2**). There was also evidence for a tradeoff – discriminability decreased as both Pair Preference scores and scoring function weights increased, and preference judgments decreased as Perceptual Difference and Name Difference increased. This finding supports our earlier claim that care must be taken to design palettes that balance both discriminability and aesthetic preference. We also found that Name Difference, not Perceptual Distance, might better predict discriminability. This would also support Demiralp et al.'s previous findings that suggested Name Difference is a better measure of color distance than Perceptual Difference [3].

Additionally, our results suggest that smaller palettes are more discriminable (**P1**), that palette size modulates discriminability and preference ratings (**P3**), and that slider configurations significantly predict behavior (**P4**). We provide additional analysis and discussion for each prediction in the Supplementary Material.

Last, differences in discriminability and aesthetic preference trends

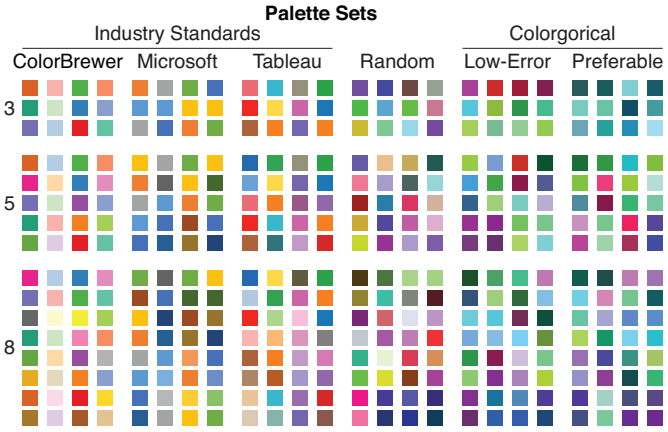


Fig. 7. Exp. 3 palettes: ColorBrewer (Dark2, Pastel1, Set1, Set2); Microsoft (all); Tableau (10, Blue-Red, Green-Orange, Purple-Gray); Colorgorical and Random palettes varied across participants.

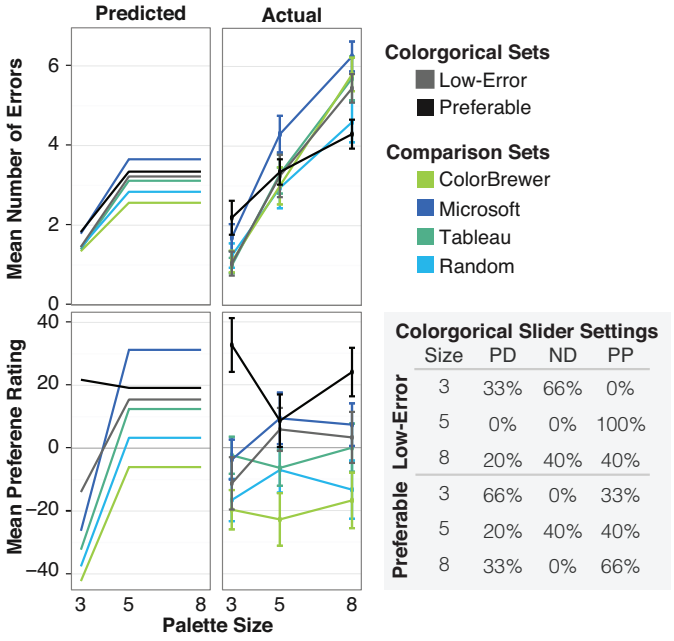


Fig. 8. Palette-behavior predictions and actual results for Experiment 2 (e.g., 4 errors = 25% error rate). Prediction models were trained on Experiment 1 palette scores and behavioral responses. Error bars show SEM. The table shows the Exp.2 Colorgorical slider settings (PD: Perceptual Distance; ND: Name Difference; PP: Pair Preference).

across palette sizes motivate additional research beyond pairwise theoretical models of color discrimination and preference rating.

7 Exp. 2: Colorgorical-Others Benchmark

Experiment 2 compares palettes generated by Colorgorical Low-Error and Preferable slider settings to commonly used "benchmark" palettes (ColorBrewer, Microsoft Excel, and Tableau; Fig. 7). We also included randomly sampled palettes with noticeably different colors to simulate palettes made by someone without design expertise who tried to choose colors that were not confusable. We predicted that:

P Colorgorical Low-Error and Preferable settings would produce palettes that are at least as discriminable and typically more preferable compared to the majority of benchmarks

We based this prediction on expected outcomes of Colorgorical and benchmark palettes by applying regressions modeled on Experiment 1 palette scores and behavioral responses to the palette scores of Experiment 2 palette sets. As shown in Figure 8, Colorgorical palettes were expected to create more preferable palettes, with the exception of Microsoft 5- and 8-color palettes, which were predicted to outperform both Colorgorical settings. We also expected that Colorgorical would produce palettes with error rates similar to Tableau across all three sizes. We specified planned comparisons to test these predictions with the human-subject data from Experiment 1.

7.1 Methods

7.1.1 Participants

75 participants (recruited through Amazon Mechanical Turk; paid $1) completed the discrimination task and 60 completed the preference task. All gave informed consent, and the Brown University IRB approved the experiment protocol. All self-reported having normal color vision. 15 discrimination participants were less than 60% accurate and were discarded, per Experiment 1 procedure (3-colors: $n = 0$, 5-colors: $n = 7$, 8-colors: $n = 8$). Participants were divided equally across size conditions ($n = 20$ per size), and there was a significant effect between discard rate and size ($\chi^2(2) = 6.878, p = 0.032$).

7.1.2 Design, Displays, & Procedure

Palette size (3,5,8) varied between subjects and the rest of the factors varied within-subject. Participants in the discriminability task completed 96 trials (6 palette sets {Colorgorical Low-Error and Preferable, ColorBrewer, Microsoft, Tableau, Random} × 4 palettes taken from each set × 4 repetitions). Participants in the preference rating task were presented with 24 trials (6 palette sets × 4 palettes, no repetition).

The benchmark palette sets included four palettes from each palette group's larger collection (Fig. 7). Microsoft palettes included all four available palettes in Microsoft Excel for Mac (v.15.8). ColorBrewer palettes included four of the eight available palettes, including those with the greatest minimum Perceptual Distance and excluding palettes with niche purposes (e.g., "Paired") [6]. Tableau palettes included the default Tableau 10 and the three palettes that were not designed for niche applications. We created random palettes by randomly sampling discriminable colors in RGB space for each participant (Sec. 4.3.1). All participants saw the same benchmark palettes aside from random. Each participant was given different random and Colorgorical palettes to test each palette type's full potential variance. The Low-Error and Preferable palettes were made with settings described at the end of Section 6 (Fig. 8). Otherwise, the design, stimuli, and procedure were the same as Experiment 1. The discrimination task took ~5 minutes to complete and the preference rating task took ~10 minutes to complete.

7.2 Results and Discussion

We focused only on error and preference rating data (not RT) because error and RT results in Experiment 1 were similar and because we chose the Colorgorical palettes based on error rates and preference ratings. All reported t-tests were paired sample and two-tailed.

We first conducted two 6 palette set (within-subject) × 3 palette size (between-subject) mixed-design ANOVAs: one for error rates (averaged over replications) and a second for preference ratings. For error, there were main effects of palette set ($F(5,285) = 3.538, p = 0.004$), palette size ($F(2,57) = 59.34, p < 0.001$), and a 2-way interaction between them ($F(10,285) = 5.896, p = 0.01$). For preference ratings, there was a main effect of palette set ($F(5,285) = 13.235, p < 0.001$) with no effect of palette size ($F(2,57) = 0.258, p = 0.773$) and no interaction ($F(10,285) = 1.283, p = 0.239$). As shown in Figure 8, error increased with size, but preference ratings were more stable as size increased. Palette set differences are shown through the vertical separation of behavioral responses across palette sets. Our planned comparisons below delve into these effects, and they largely support the trends in our predictive models based on palette score with (although size does not show the predicted effect for preference ratings).

7.2.1 Colorgorical Low-Error vs. Preferable Palettes

We first tested whether the error and preference differences between Colorgorical-Low-Error and -Preferable palettes replicated the results of Experiment 1. As in Experiment 1, the Preferable palettes were preferred to the Low-Error palettes for 3- and 8-color palettes ($t(19) = 3.573, -3.79; p = 0.002, 0.001;$), but not for 5-color palettes ($t(19) = -0.405, p = 0.690$). There were fewer errors for the 3-color Low-Error palettes than for the Preferable palettes ($t(19) = 3.286, p = 0.004$), but there was no difference for the 5-color palettes ($t(19) = 0.195, p = 0.847$). The only test that was inconsistent with our previous findings was that error rates for 8-colors were lower for Preferable palettes than for Low-Error palettes ($t(19) = 2.113, p = 0.048$). The reason for this result is unknown.

7.2.2 Comparing Colorgorical to industry standard palettes

We next tested our prediction that Colorgorical palettes would be as discriminable and typically more preferable than the benchmark palettes. The tests were planned *a priori* based on predictions from Colorgorical and benchmark palette scores described below (Fig. 8). We conducted 48 paired two-sample t-tests comparing participants' discrimination error and preference ratings within the Colorgorical palettes and between the Colorgorical palettes and the four benchmark palette sets within each palette size (Fig. 8).

Error rate. Based on the model predictions (Fig. 8), we expected that error would not significantly differ between Colorgorical Low-Error palettes and all benchmarks except for Microsoft, where we predicted that Low Error palettes would elicit fewer errors. For 5- and 8-colors we predicted that Low-Error errors would be similar to Tableau, worse than ColorBrewer and Random, and better than Microsoft. We made the same predictions for Preferable palettes, except that 3-color error might only be as good as Microsoft, and 5- and 8-color error might be worse than Tableau.

Performance for Low-Error palettes was slightly better than expected. There were significantly fewer errors for 5-color Low Error than for 5-color Microsoft ($t(19) = 2.396, p = 0.027$) and no significant difference from the other benchmarks ($t(19) < 1.628, p \geqslant 0.12$).

Colorgorical-Preferable error also matched our predictions because there was always at least one benchmark that had non-significantly different error rates compared to the setting ($t(19) \leqslant 1.898, p \geqslant 0.073$). Unexpectedly, Colorgorial-Preferable palettes led to significantly lower error than 8-color ColorBrewer and Microsoft palettes ($t(19) \geqslant 2.910, p \leqslant 0.009$). However, consistent with our predictions, 3-color Colorgorial-Preferable led to significantly more errors than ColorBrewer, Tableau, and Random benchmarks ($t(19) = 2.531, 3.644, 3.047; p = 0.020, 0.002, 0.007$, respectively).

The fewer errors for random than for Colorgorical preferable may be surprising, but it is consistent with our earlier observations. There is a high likelihood that three randomly sampled colors will be far apart in our quantized CIELAB space, leading to very high discriminability but also low preference. As the number of randomly sampled colors increases, discriminability decreases, as shown in the non-significant comparisons to 5- and 8-color Colorgorical Preferable. Although the Colorgorial-Preferable settings produced less discriminable results in some conditions (e.g., 3-color error), there was always at least one benchmark that lacked significantly different error rates.

Preference ratings. We predicted that both Low-Error and Preferable palettes would be more preferable in all comparisons except to 5- and 8-color Microsoft palettes (Fig. 8).

Low-Error was significantly more preferred than 5-color ColorBrewer and 8-color Random ($t(19) = 2.784, 2.279, p = 0.012, 0.034$, respectively) and was never significantly less preferred than the other benchmarks ($t(19) \leqslant 1.781, p \geqslant 0.091$). Colorgorial-Preferable palettes often led to significantly more preferable palettes (8 of 12, all but 5- and 8-color Microsoft, 5-color Random, and 8-color Tableau; $t(19) \geqslant 2.105, p < 0.05$).

Summary. Colorgorical Low-Error and Preferable palettes are almost always as discriminable and often more preferable than the current standard visualization-specific categorical color palettes (**P**). Low-Error palettes were sometimes more discriminable and more preferable or otherwise not significantly different than the benchmark palettes. Similarly, Preferable palettes often led to significantly higher preference ratings, and discriminability was not significantly different compared to at least one industry standard for all sizes. Thus, Colorgorical allows users without design expertise to create discriminable and preferable palettes that often do not have significantly different discriminability and that sometimes are more preferable than current pre-made standards.

8 Open research areas

We found that Colorgorical palettes, based on models of aesthetics and discriminability, can be as effective as expert-made visualization palettes and even more aesthetically preferable. These findings lead to several future research directions. First, given that color combination discriminability and preference can be inversely related, how can discriminability and preference be automatically optimized? Second, what alternatives to the current pairwise theoretical models might better predict discriminability and aesthetic preference for higher-order combinations (e.g., 5- or 8-colors)? Third, how would color preference models that diverge from figure/ground preference alter palette construction? For instance, how might Lin et al.'s preferable palette generation technique that learns from artist-generated training palettes [12] compare to palettes made with Pair Preference? Fourth, would the same results hold if hue filters are applied when constructing Colorgorical palettes? Fifth, how might Colorgorical help designers foresee palettes that might be indiscriminable given color deficiencies [28]?

9 Conclusion

We presented Colorgorical, a model-driven approach to generating categorical color palettes for information visualizations by configuring palette discriminability and preference. Colorgorical uses an iterative, semi-random-sampling procedure to generate palettes of a specified size. User-defined configurations work by changing the relative importance of Perceptual Distance, Name Difference, and Pair Preference scoring functions. Users can further customize palette creation by modifying the number of colors, by defining which hues to sample from, and by providing an existing palette to build upon.

The novelty of our approach stems from our departure from previous palette creation strategies. Whereas previous palette creation tools focused primarily on discriminability or favored color relations in harmonic templates whose empirical validity is questionable (e.g., Adobe Color [21]), Colorgorical generates palettes with user-defined relative importances for discriminability and aesthetic preference (Sec. 3). Our color sampling approach also differs in strategy from pre-made palette sets such as ColorBrewer, in which categorical palettes are generated by first choosing colors representing different names and then varying each palette color's value [1].

Empirical tests show that each of Colorgorical's sliders, which are used to balance palette discrimination and preference, measure different aspects of color (Sec. 5) and modulate behavior as they were designed to do (e.g., weighting discriminability sliders increases discriminability performance) (Sec. 6).

Empirical tests that compare Colorgorical palettes and industry standards revealed that our model-derived palettes are as effective as, and sometimes better than, current categorical color palette standards. Our findings also indicate that the number of colors may alter the effectiveness of pair-based discriminability and preference scores. Colorgorical also improves upon industry standards by giving users the flexibility to create their own discriminable and preferable palettes while enforcing visualization design constraints. These results indicate that Colorgorical provides an effective way to create categorical visualization color palettes. Colorgorical is open-sourced at `http://vrl.cs.brown.edu/color`.

Acknowledgments

This material is based upon work supported by a National Science Foundation Graduate Research Fellowship under Grant No. DGE-1058262 and Brown University's Center for Vision Research.

REFERENCES

[1] C. A. Brewer, G. W. Hatchard, and M. A. Harrower. Colorbrewer in print: A catalog of color schemes for maps. *Cartography and Geographic Information Science*, 30(1):5–32, 2003.

[2] M. E. Chevreul. *The principles of harmony and contrast of colours, and their applications to the arts*. Van Nostrand Reinhold, New York, NY, USA, 1987 (1839).

[3] Çağatay. Demiralp, M. Bernstein, and J. Heer. Learning perceptual kernels for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1933–1942, Dec 2014.

[4] C. C. Gramazio, K. B. Schloss, and D. H. Laidlaw. The relation between visualization size, grouping, and user performance. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1953–1962, Dec 2014.

[5] S. Haroz and D. Whitney. How capacity limits of attention influence information visualization effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2402–2410, Dec 2012.

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

[7] C. G. Healey. Choosing effective colours for data visualization. In *Visualization*, pages 263–270, Oct 1996.

[8] J. Heer and M. Stone. Color naming models for color selection, image editing and palette design. In *ACM Human Factors in Computing Systems (CHI)*, pages 1007–1016, New York, NY, USA, 2012. ACM.

[9] G. Hu, Z. Pan, M. Zhang, D. Chen, W. Yang, and J. Chen. An interactive method for generating harmonious color schemes. *Color Research and Application*, 39(1):70–78, 2014.

[10] J. Itten. *The art of color*. Van Nostrand Reinhold, New York, NY, USA, 1961.

[11] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer. Selecting semantically-resonant colors for data visualization. *Computer Graphics Forum*, 32(3pt4):401–410, 2013.

[12] S. Lin, D. Ritchie, M. Fisher, and P. Hanrahan. Probabilistic color-by-numbers: Suggesting pattern colorizations using factor graphs. *ACM Trans. Graph.*, 32(4):37:1–37:12, July 2013.

[13] M. R. Luo, G. Cui, and C. Li. Uniform colour spaces based on ciecam02 colour appearance model. *Color Research & Application*, 31(4):320–330, 2006.

[14] 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.

[15] Y. Matsuda. *Color Design*. Asakura Shoten, 1995.

[16] B. A. Maxwell. Visualizing geographic classifications using color. *The Cartographic Journal*, 37(2):93–99, 2000.

[17] B. J. Meier. Ace: A color expert system for user interface design. In *Proceedings of the ACM SIGGRAPH Symposium on User Interface Software*, UIST '88, pages 117–128, New York, NY, USA, 1988.

[18] B. J. Meier, A. M. Spalter, and D. B. Karelitz. Interactive color palette tools. *IEEE Computer Graphics and Applications*, 24(3):64–72, May 2004.

[19] R. Munroe. Color survey results, May 2010.

[20] A. H. Munsell. *A grammar of color*. Van Nostrand Reinhold, New York, NY, USA, 1969 (1921).

[21] P. O'Donovan, A. Agarwala, and A. Hertzmann. Color compatibility from large datasets. *ACM Trans. Graph.*, 30(4):63:1–63:12, 2011.

[22] W. Ostwald. *Colour Science (Vol. II)*. Winsor and Newton, Ltd., London, UK, 1933.

[23] L.-C. Ou, P. Chong, M. R. Luo, and C. Minchew. Additivity of colour harmony. *Color Research & Application*, 36(5):355–372, 2011.

[24] L.-C. Ou and M. R. Luo. A colour harmony model for two-colour combinations. *Color Research & Application*, 31(3):191–204, 2006.

[25] S. E. Palmer and W. S. Griscom. Accounting for taste: Individual differences in preference for harmony. *Psychonomic Bulletin & Review*, 20(3):453–461, 2012.

[26] S. E. Palmer and K. B. Schloss. An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences*, 107(19):8877–8882, 2010.

[27] R. Ratcliff. Methods for dealing with reaction time outliers. *Psychological bulletin*, 114(3):510, 1993.

[28] K. Reinecke, D. R. Flatla, and C. Brooks. Enabling designers to foresee which colors users cannot see. In *ACM Human Factors in Computing Systems (CHI)*, pages 2693–2704, New York, NY, USA, 2016. ACM.

[29] K. B. Schloss, D. Hawthorne-Madell, and S. E. Palmer. Ecological influences on individual differences in color preference. *Attention, Perception, & Psychophysics*, 77(8):2803–2816, 2015.

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

[31] V. Setlur and M. Stone. A linguistic approach to categorical color assignment for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):698–707, Jan 2016.

[32] G. Sharma, W. Wu, and E. N. Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research and Applications*, 30(1), 2005.

[33] M. Stone, D. A. Szafir, and V. Setlur. An engineering model for color difference as a function of size. In *Proceedings of the Color and Imaging Conference*, pages 228–233, 2014.

[34] F. Szabó, P. Bodrogi, and J. Schanda. Experimental modeling of colour harmony. *Color Research & Application*, 35(1):34–49, 2010.

[35] C. Taylor and A. Franklin. The relationship between color–object associations and color preference: Further investigation of ecological valence theory. *Psychonomic Bulletin & Review*, 19(2):190–197, 2012.

[36] K. Yokosawa, K. B. Schloss, M. Asano, and S. E. Palmer. Cross-cultural studies of color preferences: Us and japan. *Cognitive Science*, 2015.

[37] A. Zeileis, K. Hornik, and P. Murrell. Escaping rgbland: Selecting colors for statistical graphics. *Computational Statistics and Data Analysis*, 53(9):3259 – 3270, 2009.

Colorgorical: Creating discriminable and preferable color palettes for information visualization
Supplementary Material

Connor C. Gramazio, *Student Member, IEEE*, David H. Laidlaw, *Fellow, IEEE*, Karen B. Schloss

Abstract—.

Index Terms—Aesthetics in Visualization, Color Perception, Metrics & Benchmarks, Visual Design, Visualization

1 Overview

We present (1) additional, more thorough explanations of how each of Colorgorical's palette scores operate; (2) an analysis of how the scores are related to one another; (3) extended analysis of Experiments 1 and 2; and (4) example palettes made with 20 representative Colorgorical slider settings. We include supplementary figures and the tables presenting the statistics from our analyses.

2 Name Uniqueness and Difference score explanations

Both Name Uniqueness and Name Difference are color-term association statistics that were originally created by Heer and Stone [2]. The color-name associations map every color in a quantized 8,325-color CIELAB space to 153 popular color names, which was based on data from an XKCD crowdsourcing experiment. Name Uniqueness refers to their "name saliency" statistic, which we renamed to avoid confusion with color saliency.

Name Difference can be thought of as how much two colors' association mappings overlap, whereas Name Uniqueness can be thought of as how uniformly distributed a colors' associations are to the 153 names. Each scoring function is illustrated in Supplementary Figure 1.

3 Interpolating the Pair Preference *Coolness* term

In Supplementary Figure 2 we show the linear interpolation results that calculate coolness values for CIE LCh space. The interpolated values are derived from the coolness values of the 32 Munsell colors used in the original Schloss and Palmer pair preference in-lab experiment [4]. These values were calculated by counting the number of steps each of the colors was from the color 10R in Munsell color space.

4 Runtime performance

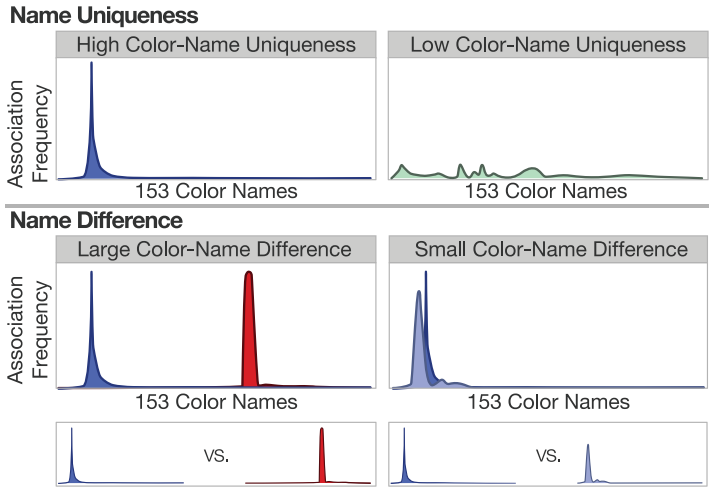
To evaluate model runtime with respect to palette size, we profiled single-palette generation with 1 to 20 colors 50 times each on a Mid 2012 MacBook Pro Retina with a 2.6 GHz Intel Core i7 CPU and 16GB 1600MHz DDR3 RAM. Average initialization time was 0.14 seconds (SEM = 0.004). If a palette was returned before reaching the required number of colors, it was discarded and the test was run again. Runtime performance increased linearly in the number of colors (S.Fig. 3).

5 Palette Score Evaluation

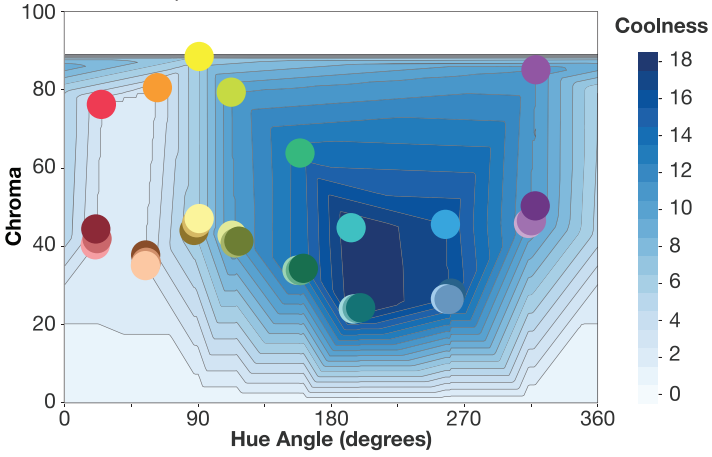
The aim of this experiment was to assess whether the model could be simplified by removing redundant scoring functions. For instance, if Perceptual Distance were to explain most of the variance in Name Difference scores, then the Name Difference scoring function could be removed from the model with little effect on palette output.

To examine how similar the four Colorgorical scoring functions were to one another, we tested the degree of independence between *palette scores*.

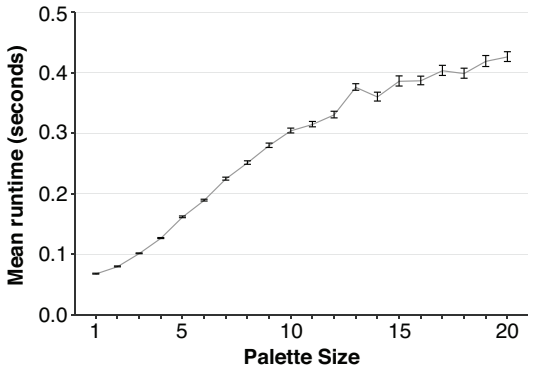
- *C. Gramazio and D. Laidlaw are with the Dept. of Computer Science at Brown University. E-mail: {connor,dhl}@cs.brown.edu.*
- *K. Schloss is with the Dept. of Cognitive, Linguistic, and Psychological Sciences at Brown University. E-mail: karenschloss@gmail.com.*



S.Fig. 1. Name Uniqueness and Name Difference. High Name Uniqueness scores are focally distributed color-name associations, whereas low Name Uniqueness scores are more uniform distributions. Name Difference scores are proportional to the difference between color-name association distributions. The blue and red example is large because there is little color-name association overlap.



S.Fig. 2. Interpolated coolness layered on chroma and hue from CIE LCh. The original Schloss & Palmer coolness values are derived from how many steps each color used in the experiment used to derive Pair Preference is from the color 10R in Munsell color space.



S.Fig. 3. Runtime performance of Colorgorical for 20 palette sizes. Error bars show standard error for each number of colors' 50 tests.

Palette scores are derived by taking the minimum scoring function output given all color pairs in a palette. We use the minimum score based on our model's assumption that a palette is only as preferable or discriminable as its lowest pair.

We also tested (1) how the three discriminability palette scores compare to Pair Preference, and (2) how each of the four palette scores changes with the number of colors in a palette. We predicted that:

P1 All palette scores measure different color-relation information and are not redundant

P2 Pair Preference would be a negative predictor of the Perceptual Distance and Name Difference

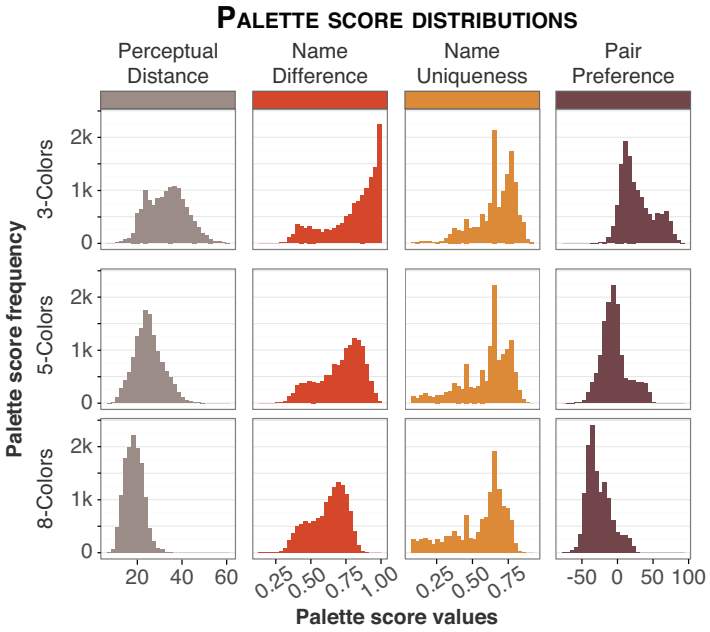
5.1 Methods

To test the full range of Colorgorical output, we created a representative set of 39,600 palettes. This collection was made using 66 unique slider settings, which tested different relative importance of scoring functions, and palettes of 3, 5, and 8 colors. The 66 settings are the different unique combinations a user could make by dragging each slider to 0%, 50%, or 100%. The combinations ignore duplicate settings encountered when moving one or more sliders to 0%. For instance, if three sliders are turned to 0%, any non-0% position of the fourth slider would give it a relative importance of 100%.

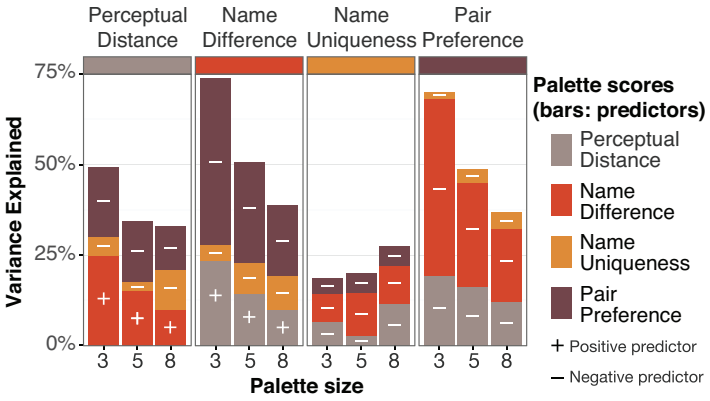
5.2 Results & Discussion

Before testing the relation between each of the four palette scores, we first plotted the distribution for each of the palette scores across the different palette sizes (S.Fig. 4). One noticeable trend is that the palette scores decrease as a whole with respect to palette size. To test whether this trend was significant we correlated each collection of palette scores with palette size ({3,5,8}-colors) and found that each trend was significant (Pearson's $r(39598) = -0.667(\text{PD}), -0.429(\text{ND}), -0.267(\text{NU}), -0.727(\text{PP}); p < 0.001$). These trends might originate from a combination of two sources. First, increasing the number of colors leaves successively fewer regions of available color pace to sample from. Second, using wider swaths of color space increases the likelihood that there is a low score in the exponentially growing number of color pairs in a palette. Using different aggregation techniques (e.g., leave-lowest-out or averaging) might result in higher palette scores, but would also go against our assumption that a palette is only as discriminable or preferable as the worst-performing pair of colors in a palette. However, we believe that testing this assumption would be an interesting direction for future research.

After, we tested whether Perceptual Distance, Name Difference, Name Uniqueness, and Pair Preference *palette scores* were independent (P1). To evaluate this prediction we used multiple linear regression analyses to predict palette scores as a function of the other three (e.g., predicting Perceptual Distance with Name Difference, Name Uniqueness, and Pair Preference). We conducted separate regressions for each palette size, resulting in 12 regressions (4 palette scores $\times$ 3 sizes) used to predict 13,200 palettes. For each regression, all predictors explained a significant amount of variance (all $F(3, 13196) \geqslant 1007.676$, all $p < 0.001$; all $t(13196) \geqslant 16.26$, all $p < 0.001$).



S.Fig. 4. Palette score distributions for each of the 39,600 palettes used in the palette score verification and Experiment 1.



S.Fig. 5. Percent of explained variance of 12 linear regressions that test whether palette scores predict one another (stack height). Each bar shows the percent of explained variance for each palette score. All regressions and predictors were significant.

Supplementary Figure 5 shows the relative importance of the predictors in each regression model [1]. Perceptual Distance and Name Difference showed similar trends in that both were positive predictors of one another, Name Uniqueness was a small negative predictor, and Pair Preference was a large negative predictor. The largest difference between Perceptual Distance and Name Difference was that Pair Preference explained a much larger portion of Name Difference's variance (and vice versa; 3-Color ND predicted by PP = 46.1%; 3-Color PP predicted by ND = 48.9%). This strong negative association could be linked to the hue similarity term in Pair Preference, which might sometimes create a discrimination-preference trade-off (P2). These findings suggest that Pair Preference is more strongly related to the Name Difference of colors than to Perceptual Distance.

We concluded from these results that each function measured sufficiently different color information (P1) because a single palette score never predicted greater than 50% of variance in another. Therefore, we kept all of the four scores in the model while generating the stimuli for Experiment 1.

6 Exp. 1: palette score correlation with behavior

6.1 Selecting palette score binning widths

A large problem when correlating palette scores with behavior is the individual differences that occur between subjects. Another problem is that near-scores are treated as separate values, leading to uninformative correlations. To avoid both problems, we quantized each palette score into 15 bins. Bin widths were calculated using the full range values over 3-, 5-, and 8-colors for each score. Our goal when selecting the number of bins to use was to maximize the number of subjects who were shown all bins. In other words, we wanted to avoid having bins with few subjects in them to improve the consistency of analyses.

We first attempted binning using the Freedman-Dianconis rule, which resulted in 29 bins:

$$\text{number of bins} = \left\lceil \frac{\max(x) - \min(x)}{h} \right\rceil \tag{1}$$

$$h_{\text{F-D}} = 2\frac{\text{IQR}(x)}{n^{1/3}} \tag{2}$$

We also attempted binning using Sturge's formula, which resulted in 13 bins:

$$h_{\text{Sturge}} = \lceil \log_2 n + 1 \rceil \tag{3}$$

Our ultimate selection method relied on picking bins after charting different widths because the Freedman-Diaconis rule resulted in many bins with few subjects for some sizes, and Sturge's formula was too coarse of a score description. We tested 10 to 30 bins in increments of 5, reflecting the range between rounded Sturge and Freedman-Dianconis bin suggestions. The chart is shown in Supplementary Figure 6.

6.2 Correlation results

Binned palette score correlation results are shown in Supplementary Figure 8 (Pearson's r). An alternative to our binning approach is to correlate within-subject unbinned palette scores with behavioral measures, and then apply Fisher's Z transform to average Pearson's r between subjects[1]. A side-by-side comparison of the correlation results for each method is shown in Supplementary Figure 7. To test for significance in the individual correlations, we conducted one-sample t-tests for each palette score within each size to compare the mean of the individual subjects' correlations against zero. The magnitude of Pearson's r was smaller for the mean of the individual correlations (S.Fig. 7, left) compared to the correlations across means between the mean data (S.Fig. 7, right). The difference in magnitude is expected, given that averaging between subjects reduces the noisy variance stemming from individual differences. Nonetheless, the pattern of results is similar: 8 bars (of 36) are significant with binning that are not with Fisher's transform, and 4 bars are significant with Fisher's transform that are not with binning (12 of 36 total). Both methods show few significant Name Uniqueness correlations despite the number of false negatives and false positives. Therefore, both correlation methods support our decision to remove Name Uniqueness from further analysis.

6.3 Slider settings' mapping onto behavior

To capture how manipulating sliders' relative importances mapped onto the behavioral data captured in Experiment 1, we created a set of Barycentric plots (S.Fig. 9). Each facet of the plot shows a different size $\times$ behavioral measure condition, and each circle is one of the 20 tested slider settings. The triangle fill colors are the Barycentric interpolated values between each tested setting and a thicker stroke indicates the Experiment 2 Low-Error and Preferable palette settings. Of note, Perceptual Distance was more amenable to preserving preference ratings when increased. The neutral preference rating trend across 5-color slider settings discussed in the primary manuscript is also reflected, as is the 8-color response time drop off compared to 3- and 5-color response times.

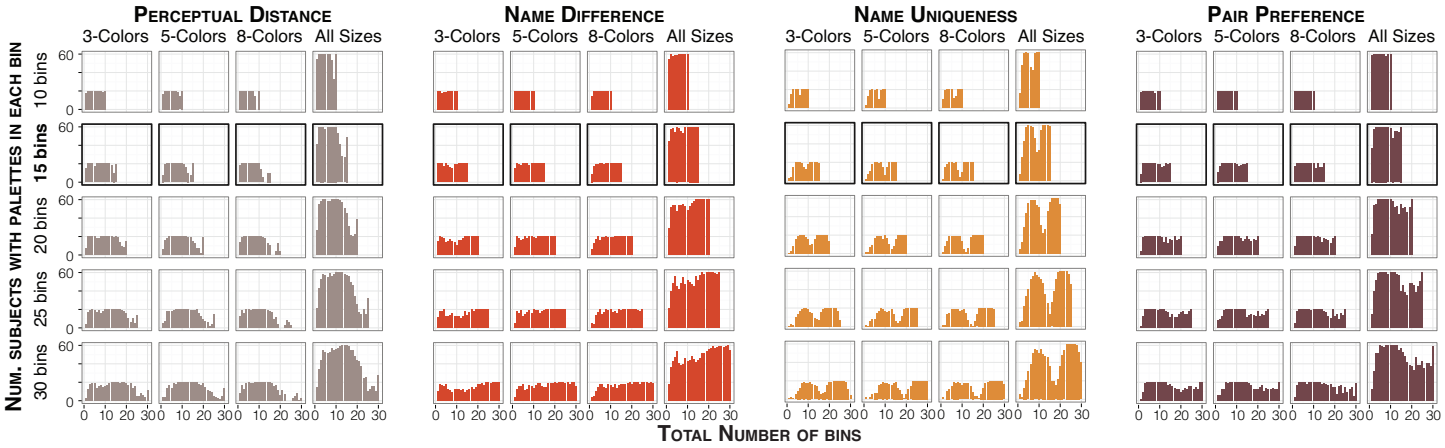
7 Exp. 2: Supplementary Material

The full list of palette set means and standard errors for the Experiment 2 palettes are listed in S.Table 1.

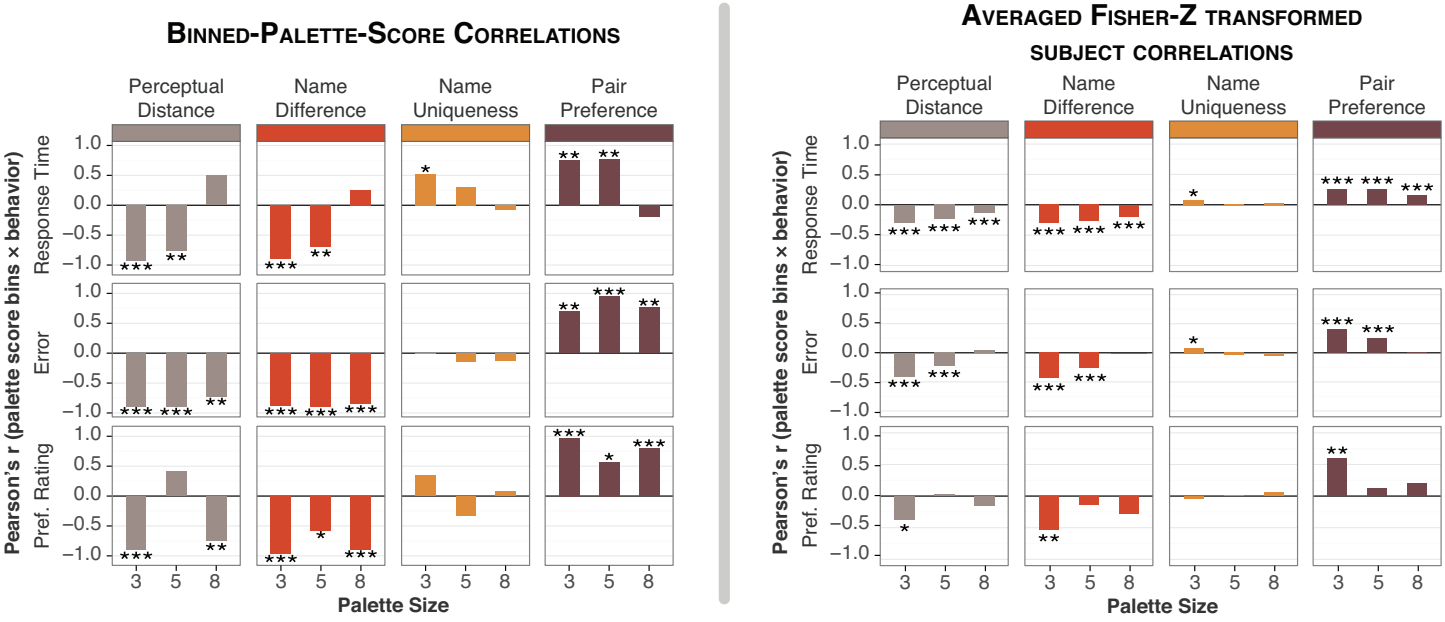
[1] Fisher's Z transform converts correlation coefficient distributions to be more normal-like [3]

Size	Palette Set	Measure	Mean	Std. Error
3	ColorBrewer	Error	1.1	0.27
3	Low-Error	Error	1.05	0.303
3	Preferable	Error	2.2	0.427
3	Microsoft	Error	1.7	0.341
3	Random	Error	1.25	0.307
3	Tableau	Error	1	0.192
5	ColorBrewer	Error	3	0.465
5	Low-Error	Error	3.25	0.532
5	Preferable	Error	3.35	0.319
5	Microsoft	Error	4.3	0.459
5	Random	Error	2.95	0.51
5	Tableau	Error	3.3	0.493
8	ColorBrewer	Error	5.8	0.421
8	Low-Error	Error	5.45	0.359
8	Preferable	Error	4.3	0.363
8	Microsoft	Error	6.25	0.376
8	Random	Error	4.6	0.505
8	Tableau	Error	5.7	0.548
3	ColorBrewer	RT	1474.66	78.458
3	Low-Error	RT	1479.618	79.75
3	Preferable	RT	1542.629	87.427
3	Microsoft	RT	1568.527	94.217
3	Random	RT	1454.229	79.477
3	Tableau	RT	1453.156	81.904
5	ColorBrewer	RT	1824.114	81.677
5	Low-Error	RT	1875.154	83.546
5	Preferable	RT	1858.203	79.526
5	Microsoft	RT	1955.883	88.54
5	Random	RT	1837.006	70.413
5	Tableau	RT	1865.575	75.587
8	ColorBrewer	RT	1769.46	78.733
8	Low-Error	RT	1780.639	73.004
8	Preferable	RT	1772.981	68.538
8	Microsoft	RT	1755.502	96.269
8	Random	RT	1719.034	77.138
8	Tableau	RT	1770.159	87.515
3	ColorBrewer	Pref. Rating	-19.744	6.249
3	Low-Error	Pref. Rating	-11.369	8.321
3	Preferable	Pref. Rating	32.644	8.56
3	Microsoft	Pref. Rating	-3.806	6.352
3	Random	Pref. Rating	-16.663	6.724
3	Tableau	Pref. Rating	-2.444	5.897
5	ColorBrewer	Pref. Rating	-22.837	8.374
5	Low-Error	Pref. Rating	5.787	6.845
5	Preferable	Pref. Rating	8.55	8.364
5	Microsoft	Pref. Rating	9.312	8.181
5	Random	Pref. Rating	-7.1	7.205
5	Tableau	Pref. Rating	-6.45	5.692
8	ColorBrewer	Pref. Rating	-16.788	8.818
8	Low-Error	Pref. Rating	3.25	8.119
8	Preferable	Pref. Rating	24.038	7.715
8	Microsoft	Pref. Rating	7.312	6.776
8	Random	Pref. Rating	-13.375	9.262
8	Tableau	Pref. Rating	-0.012	7.65

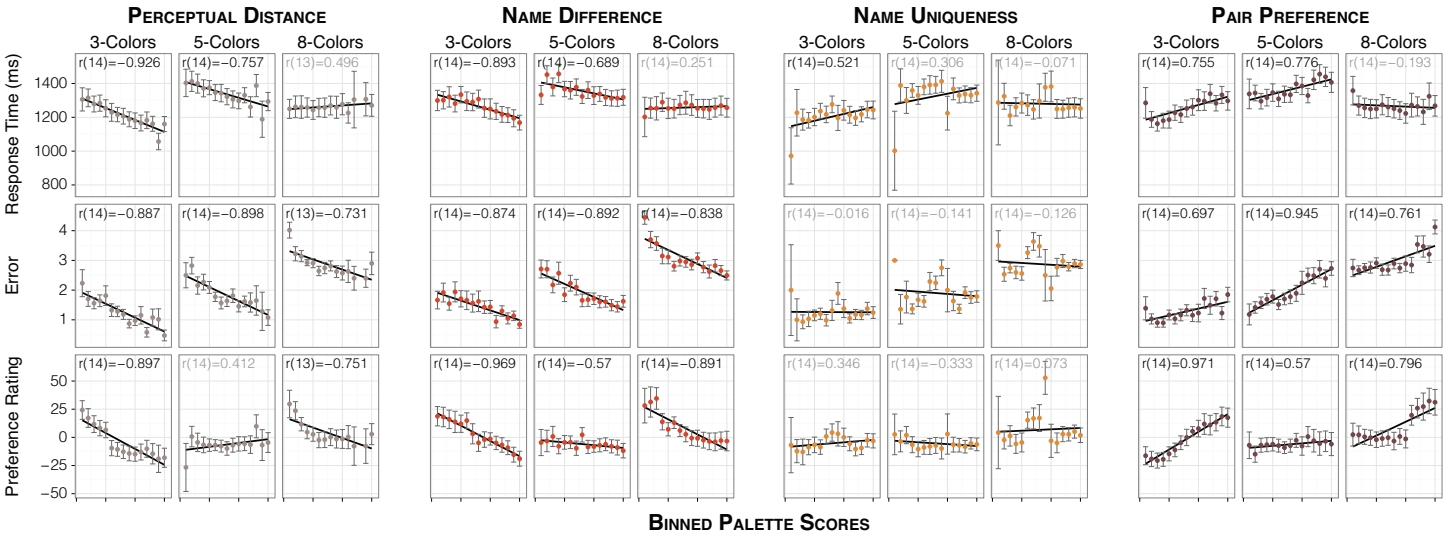
S.Table 1. The mean and standard error responses for each palette set $\times$ size combination. Low-Error and Preferable palette sets are the two Colorgorical settings included in Experiment 2.



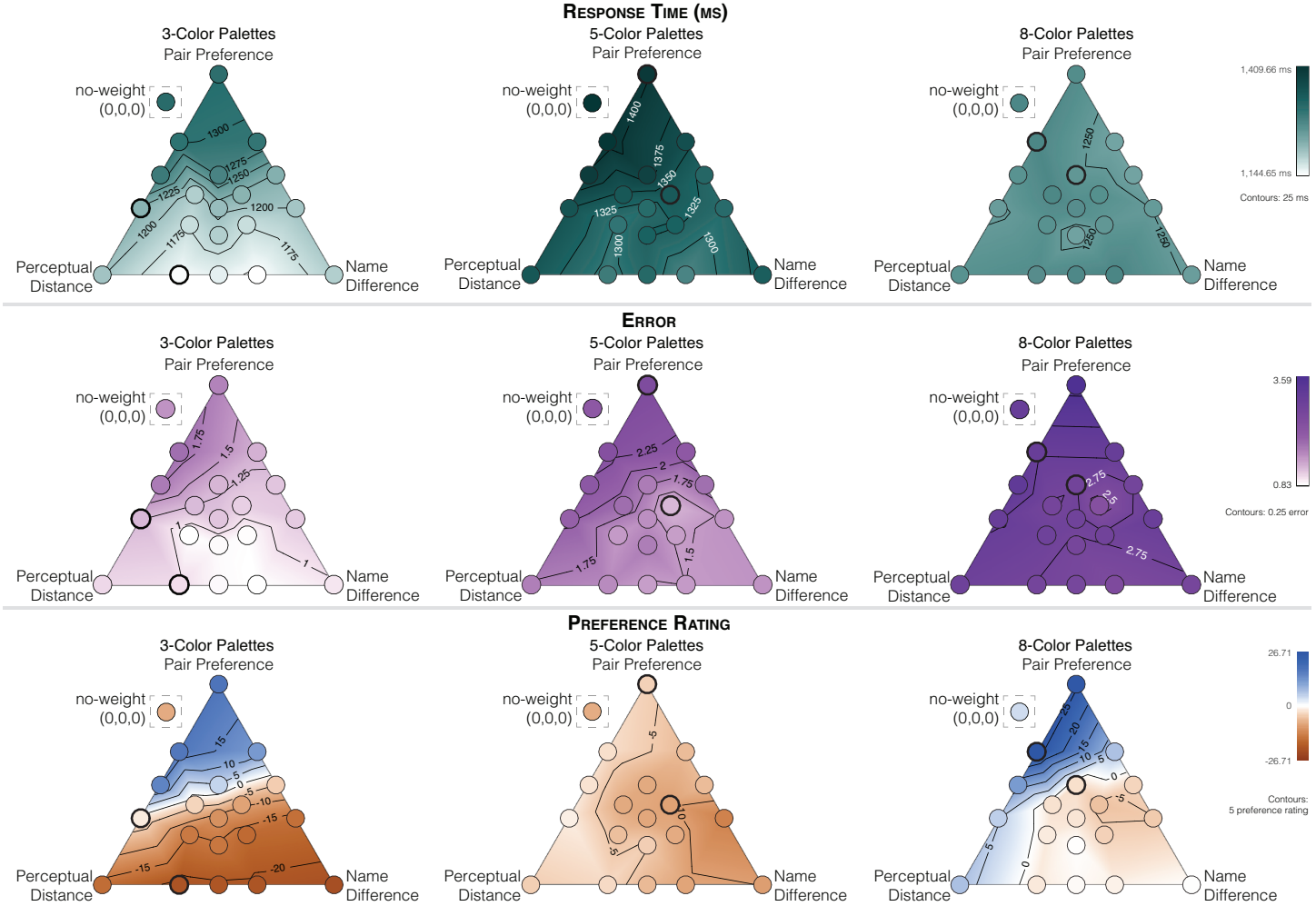
S.Fig. 6. Plots of different bin amounts for Experiment 1, where each bar represents the number of subjects who were shown palettes within a given bin. An ideal setting first minimizes the number of low-subject bins, and then favors a larger number of bins to better describe the distribution.



S.Fig. 7. Side-by-side comparison of correlation results. The left chart is reproduced from the primary manuscript and shows the correlations for each palette score × size condition using 15 bins. The right chart shows the same conditions, but the correlations are the average correlations between subjects using Fisher's Z transform.



S.Fig. 8. Results from Exp.1 correlations between palette scores and behavioral measures for 3-, 5-, and 8-color palettes. To compare similar palette scores, each score was quantized into 15 bins. Error bars show standard error within each bin. Pearson's correlation coefficients are shown in black ($p < 0.05$) or grayed out ($p \geqslant 0.05$) text. Lines show a linear regression fit between palette score and behavior.



S.Fig. 9. Results from Experiment 1 investigating how slider settings for each scoring function (i.e., relative importance) change discrimination performance and preference ratings. Colored circles represent the settings tested. Error refers to the total number of errors made with a particular slider setting, where each of the original 66 slider settings had 10 repetitions within each subject (i.e., 1 error = 10%).

Size	Measure	$t(19)$	p
3	Pref. Rating	-3.573	0.002
5	Pref. Rating	-0.405	0.69
8	Pref. Rating	-3.79	0.001
3	Error	-3.286	0.004
5	Error	-0.195	0.847
8	Error	2.113	0.048
3	RT	-1.513	0.147
5	RT	0.309	0.761
8	RT	0.221	0.828

S.Table 2. Experiment 2 t-tests between Colorgorical Low-Error and Preferable settings. Negative t-values favor Preferable.

8 Linear regression and t-test analysis results

The tables for Palette Verification and Experiments 1 and 2 linear regressions and t-tests are shown in the tables below (S.Tables 2-5). All t-tests were paired and two-tailed.

9 Colorgorical output examples

Below are examples of the 20 slider settings that can be made by changing Perceptual Distance (PD), Name Difference (ND), and Pair Preference (PP) sliders to 0%, 50%, and 100%. We left Name Uniqueness at 0% given Experiment 1 results. Note that color appearance is slightly off, given that Colorgorical designs RGB palettes (figures in both this document and in the primary manuscript are rendered in CYMK). As such, we include palette color D65 CIELAB coordinates to the right of all palettes.

Measure	Colors	Setting	Benchmark	$t(19)$	p
Error	3	Low-Error	ColorBrewer	-0.139	0.891
Error	3	Low-Error	Microsoft	-1.628	0.12
Error	3	Low-Error	Tableau	0.181	0.858
Error	3	Low-Error	Random	-0.525	0.606
Error	5	Low-Error	ColorBrewer	0.665	0.514
Error	5	Low-Error	Microsoft	-2.396	0.027
Error	5	Low-Error	Tableau	-0.103	0.919
Error	5	Low-Error	Random	0.603	0.554
Error	8	Low-Error	ColorBrewer	-0.649	0.524
Error	8	Low-Error	Microsoft	-1.417	0.173
Error	8	Low-Error	Tableau	-0.366	0.719
Error	8	Low-Error	Random	1.342	0.196
Error	3	Preferable	ColorBrewer	2.531	0.02
Error	3	Preferable	Microsoft	1.097	0.287
Error	3	Preferable	Tableau	3.644	0.002
Error	3	Preferable	Random	3.047	0.007
Error	5	Preferable	ColorBrewer	0.78	0.445
Error	5	Preferable	Microsoft	-1.727	0.1
Error	5	Preferable	Tableau	0.101	0.921
Error	5	Preferable	Random	0.867	0.397
Error	8	Preferable	ColorBrewer	-2.91	0.009
Error	8	Preferable	Microsoft	-3.456	0.003
Error	8	Preferable	Tableau	-1.898	0.073
Error	8	Preferable	Random	-0.501	0.622
RT	3	Low-Error	ColorBrewer	0.163	0.872
RT	3	Low-Error	Microsoft	-2.935	0.008
RT	3	Low-Error	Tableau	1.004	0.328
RT	3	Low-Error	Random	0.825	0.42
RT	5	Low-Error	ColorBrewer	1.311	0.205
RT	5	Low-Error	Microsoft	-2.147	0.045
RT	5	Low-Error	Tableau	0.27	0.79
RT	5	Low-Error	Random	0.594	0.559
RT	8	Low-Error	ColorBrewer	0.309	0.761
RT	8	Low-Error	Microsoft	0.457	0.653
RT	8	Low-Error	Tableau	0.223	0.826
RT	8	Low-Error	Random	1.639	0.118
RT	3	Preferable	ColorBrewer	2.008	0.059
RT	3	Preferable	Microsoft	-0.536	0.598
RT	3	Preferable	Tableau	2.171	0.043
RT	3	Preferable	Random	2.009	0.059
RT	5	Preferable	ColorBrewer	0.779	0.446
RT	5	Preferable	Microsoft	-1.804	0.087
RT	5	Preferable	Tableau	-0.156	0.878
RT	5	Preferable	Random	0.312	0.758
RT	8	Preferable	ColorBrewer	0.105	0.918
RT	8	Preferable	Microsoft	0.349	0.731
RT	8	Preferable	Tableau	0.063	0.95
RT	8	Preferable	Random	1.761	0.094
Pref. Rating	3	Low-Error	ColorBrewer	1.075	0.296
Pref. Rating	3	Low-Error	Microsoft	-0.697	0.494
Pref. Rating	3	Low-Error	Tableau	-1.193	0.248
Pref. Rating	3	Low-Error	Random	0.629	0.537
Pref. Rating	5	Low-Error	ColorBrewer	2.784	0.012
Pref. Rating	5	Low-Error	Microsoft	-0.385	0.705
Pref. Rating	5	Low-Error	Tableau	1.687	0.108
Pref. Rating	5	Low-Error	Random	1.781	0.091
Pref. Rating	8	Low-Error	ColorBrewer	1.768	0.093
Pref. Rating	8	Low-Error	Microsoft	-0.372	0.714
Pref. Rating	8	Low-Error	Tableau	0.301	0.767
Pref. Rating	8	Low-Error	Random	2.279	0.034
Pref. Rating	3	Preferable	ColorBrewer	4.439	< 0.001
Pref. Rating	3	Preferable	Microsoft	3.375	0.003
Pref. Rating	3	Preferable	Tableau	3.252	0.004
Pref. Rating	3	Preferable	Random	4.416	< 0.001
Pref. Rating	5	Preferable	ColorBrewer	3.434	0.003
Pref. Rating	5	Preferable	Microsoft	-0.064	0.95
Pref. Rating	5	Preferable	Tableau	2.105	0.049
Pref. Rating	5	Preferable	Random	1.821	0.084
Pref. Rating	8	Preferable	ColorBrewer	3.072	0.006
Pref. Rating	8	Preferable	Microsoft	1.581	0.13
Pref. Rating	8	Preferable	Tableau	2.04	0.056
Pref. Rating	8	Preferable	Random	5.2	< 0.001

S.Table 3. Experiment 2 t-tests between Colorgorical and benchmarks.

Slider settings:: PD:0.0 ND:0.0 NU:0.0 PP:0.0

[85,-45,0]; [40,40,-45]; [45,-25,-5]
[85,-45,60]; [80,-15,-5]; [50,0,-5]
[85,-45,0]; [35,-5,-5]; [80,0,65]
[70,-45,0]; [45,60,-60]; [55,10,-35]
[85,-30,0]; [80,0,-15]; [45,70,-30]

[85,-60,45]; [45,15,-55]; [65,-20,-25]; [30,15,-10]; [80,-15,20]
[85,-15,0]; [55,40,5]; [80,-20,70]; [35,5,35]; [60,0,-30]
[55,-15,-30]; [80,0,15]; [40,25,10]; [65,-55,40]; [85,-40,-15]
[40,-15,-15]; [85,-25,50]; [75,40,-15]; [50,80,-15]; [55,-50,50]
[85,-30,30]; [70,5,-15]; [30,0,20]; [85,5,40]; [60,45,-15]

[55,-15,-30]; [40,10,45]; [85,-25,75]; [75,35,15]; [40,-25,25]; [40,55,10]; [45,15,-45]; [80,-35,-5]
[85,-30,15]; [70,20,60]; [45,10,50]; [70,-45,60]; [35,45,-30]; [45,-15,-5]; [75,-20,-35]; [65,15,-20]
[70,-30,0]; [85,-10,50]; [45,65,50]; [40,15,35]; [35,45,-45]; [70,-60,35]; [75,5,-20]; [75,55,-35]
[85,-30,15]; [35,10,-15]; [60,45,30]; [60,-25,10]; [80,10,70]; [65,-60,50]; [35,55,-25]; [65,-5,-40]
[55,-15,-15]; [70,-30,60]; [75,40,-25]; [60,55,55]; [50,15,-25]; [40,55,-35]; [50,10,30]; [35,40,40]

Slider settings:: PD:0.0 ND:0.0 NU:0.0 PP:1.0

[55,-15,-15]; [80,-25,-25]; [30,-10,-20]
[70,-30,0]; [45,-25,-5]; [85,-60,25]
[85,-60,15]; [30,-15,-5]; [80,-30,-10]
[55,-15,-30]; [80,-40,-15]; [30,-20,0]
[55,-45,15]; [85,-45,15]; [35,-15,-5]

[55,-45,15]; [30,-35,25]; [80,-50,50]; [55,-15,-10]; [85,-40,-5]
[85,-45,60]; [50,-25,35]; [75,-35,15]; [30,-15,-10]; [70,-25,-30]
[70,-60,30]; [35,-35,30]; [80,-30,20]; [40,-20,-15]; [75,-15,-25]
[40,-15,-15]; [85,-10,-20]; [35,30,-60]; [60,30,-40]; [85,-40,-10]
[55,-15,-30]; [85,-30,-10]; [30,-15,-10]; [85,-40,60]; [50,-50,40]

[70,-30,-15]; [30,-20,-5]; [85,-40,30]; [50,-45,45]; [85,-70,45]; [30,15,-35]; [65,40,-45]; [75,10,-30]
[55,-15,-30]; [80,-10,-15]; [30,0,-30]; [60,20,-50]; [35,55,-60]; [80,-50,0]; [50,-35,15]; [80,-60,45]
[85,-45,45]; [50,-30,-5]; [85,-45,-10]; [45,0,-30]; [85,0,-15]; [35,50,-60]; [60,40,-25]; [30,55,-5]
[70,-15,-30]; [35,0,-35]; [80,-45,5]; [30,-20,15]; [75,-50,60]; [50,-50,45]; [80,-15,20]; [55,-20,40]
[55,-15,-30]; [85,0,-20]; [30,30,-55]; [75,40,-25]; [45,50,-10]; [60,80,-20]; [30,20,-10]; [80,-45,5]

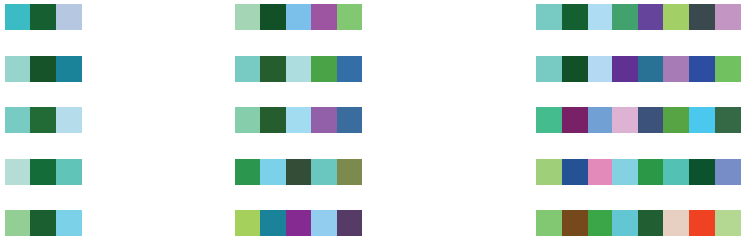
Slider settings:: PD:0.0 ND:1.0 NU:0.0 PP:0.0

[85,-45,30]; [75,-10,-20]; [40,40,-35]
[85,-15,0]; [60,-25,50]; [50,-10,-5]
[70,-60,30]; [40,50,-60]; [70,-30,-5]
[40,-15,0]; [70,30,50]; [75,-60,25]
[70,-30,0]; [30,50,-15]; [80,-65,50]

[70,-30,-15]; [75,30,-5]; [40,15,-45]; [55,85,-60]; [30,-15,-10]
[85,-60,45]; [60,10,-20]; [65,65,-10]; [80,-30,50]; [30,40,-40]
[85,-45,60]; [35,50,-25]; [80,-10,10]; [60,80,-20]; [85,-45,0]
[70,-60,30]; [40,35,-35]; [65,-5,10]; [55,25,40]; [30,-5,-5]
[85,-30,-15]; [45,25,0]; [50,-30,40]; [50,55,15]; [75,25,60]

[55,-15,-15]; [45,50,-60]; [40,55,50]; [65,-30,45]; [60,15,40]; [60,70,0]; [85,5,-20]; [85,-65,80]
[85,-30,30]; [80,25,-20]; [60,-10,5]; [45,50,-50]; [30,30,15]; [70,60,-30]; [30,-10,-10]; [55,45,55]
[85,-30,-15]; [40,20,0]; [55,-15,-15]; [35,60,5]; [50,30,55]; [65,-60,65]; [80,25,0]; [70,-25,40]
[55,-15,-15]; [80,-35,75]; [70,30,5]; [70,-65,60]; [30,15,40]; [80,-35,-15]; [35,50,-25]; [60,-15,30]
[55,-45,15]; [40,55,35]; [40,20,-35]; [85,-35,-10]; [30,55,-65]; [30,-20,15]; [80,15,25]; [60,-15,-10]

Slider settings:: PD:0.0 ND:1.0 NU:0.0 PP:1.0



[70,-30,-15]; [35,-35,25]; [80,0,-15]
[85,-30,0]; [30,-35,30]; [50,-20,-20]
[85,-45,0]; [40,-35,30]; [85,-10,-10]
[85,-15,0]; [40,-40,30]; [80,-45,0]
[85,-45,30]; [35,-35,30]; [85,-30,-20]

[85,-30,15]; [30,-30,20]; [75,-10,-30]; [50,60,-45]; [85,-60,45]
[85,-45,0]; [35,-30,25]; [85,-15,-5]; [60,-45,40]; [45,0,-35]
[85,-45,15]; [35,-30,25]; [85,-15,-15]; [50,40,-40]; [45,0,-30]
[55,-45,30]; [80,-20,-20]; [30,-15,10]; [85,-50,0]; [55,-15,30]
[85,-45,60]; [50,-20,-20]; [35,60,-45]; [80,-10,-25]; [30,20,-20]

[85,-45,0]; [35,-40,35]; [85,-10,-15]; [60,-40,20]; [35,50,-60]; [85,-45,55]; [30,-5,-5]; [70,40,-35]
[85,-45,0]; [30,-30,20]; [85,-5,-20]; [30,50,-55]; [45,-10,-25]; [60,40,-40]; [35,20,-50]; [85,-70,50]
[70,-45,15]; [30,45,-20]; [65,0,-35]; [80,35,-25]; [35,5,-25]; [60,-45,60]; [80,-30,-30]; [40,-25,15]
[85,-45,45]; [35,10,-40]; [70,50,-20]; [80,-20,-15]; [55,-50,45]; [80,-50,0]; [30,-30,15]; [60,10,-40]
[85,-60,45]; [35,15,35]; [60,-60,40]; [75,-25,-15]; [35,-30,20]; [85,5,10]; [55,65,60]; [85,-30,35]

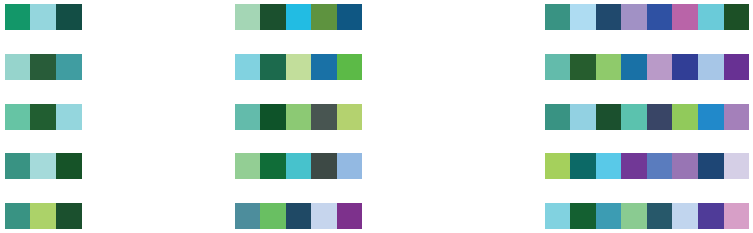
Slider settings:: PD:0.0 ND:1.0 NU:0.0 PP:0.5

[70,-30,-15]; [80,-40,45]; [45,-5,-25]
[85,-45,15]; [40,55,-55]; [85,-15,-20]
[85,-45,60]; [35,0,-5]; [65,-50,15]
[85,-30,15]; [35,55,-10]; [55,20,-70]
[85,-45,45]; [50,15,-50]; [40,60,-30]

[70,0,-45]; [35,65,-45]; [80,-30,-30]; [60,-50,55]; [35,25,-45]
[85,-30,30]; [35,55,-50]; [65,-5,-40]; [80,-40,75]; [35,10,-20]
[70,-15,-30]; [50,70,-50]; [70,-50,20]; [35,35,-25]; [40,5,-45]
[85,-30,0]; [30,45,-45]; [50,-20,-5]; [70,40,-20]; [80,0,-30]
[85,-15,0]; [30,25,30]; [75,-40,60]; [50,60,60]; [60,-20,-5]

[55,-45,15]; [35,60,-45]; [65,-5,-40]; [45,25,-25]; [85,-45,-10]; [30,-35,30]; [55,30,-70]; [70,40,-20]
[40,0,-30]; [55,40,-25]; [35,40,-75]; [55,-50,35]; [85,-10,-5]; [30,-25,15]; [85,-50,0]; [35,10,25]
[85,-15,0]; [50,-45,40]; [45,-5,-25]; [85,-65,45]; [30,45,-30]; [70,5,-45]; [60,50,-15]; [30,60,-85]
[85,-60,30]; [75,25,60]; [35,-15,5]; [80,-40,-15]; [35,20,35]; [60,-35,45]; [70,0,-30]; [50,35,-25]
[85,-60,30]; [60,35,-35]; [70,-25,30]; [55,-5,-20]; [35,55,-20]; [85,-35,-20]; [40,-35,20]; [80,5,-30]

Slider settings:: PD:0.0 ND:0.5 NU:0.0 PP:1.0



[55,-45,15]; [85,-25,-10]; [30,-20,0]
[85,-30,0]; [35,-25,15]; [60,-25,-10]
[85,-60,15]; [35,-30,20]; [85,-25,-10]
[55,-30,0]; [85,-20,-5]; [30,-35,30]
[55,-30,0]; [85,-40,55]; [30,-25,15]

[85,-30,15]; [30,-25,15]; [70,-25,-30]; [55,-35,50]; [35,-5,-30]
[85,-30,-15]; [40,-30,10]; [85,-20,30]; [45,-5,-35]; [70,-55,55]
[70,-30,0]; [30,-35,25]; [85,-55,45]; [35,-5,0]; [80,-25,45]
[85,-45,30]; [40,-40,25]; [75,-35,-15]; [30,-5,0]; [75,0,-35]
[55,-15,-15]; [75,-55,45]; [30,-5,-20]; [85,0,-15]; [35,45,-35]

[55,-30,0]; [85,-10,-15]; [30,0,-25]; [65,25,-40]; [35,30,-65]; [60,70,-45]; [85,-40,-15]; [30,-30,25]
[70,-30,0]; [35,-30,25]; [85,-55,50]; [45,-5,-35]; [70,30,-35]; [30,25,-50]; [80,0,-30]; [30,60,-60]
[55,-30,0]; [80,-15,-15]; [30,-25,15]; [80,-50,5]; [30,5,-20]; [80,-45,55]; [55,-5,-45]; [60,30,-35]
[85,-45,60]; [40,-25,-5]; [80,-30,-25]; [35,45,-45]; [55,15,-55]; [60,50,-55]; [30,5,-30]; [85,10,-20]
[85,-30,-15]; [35,-40,35]; [60,-20,-20]; [85,-50,30]; [35,-10,-15]; [85,0,-20]; [30,45,-60]; [75,45,-30]

Slider settings:: PD:1.0 ND:0.0 NU:0.0 PP:0.0

[85,-45,15]; [30,45,5]; [85,-55,60]
[85,-30,0]; [30,50,35]; [80,-65,30]
[85,-15,-15]; [30,20,0]; [75,-65,45]
[85,-15,-15]; [30,35,15]; [80,-55,60]
[55,-15,-30]; [85,-35,55]; [35,40,-60]

[40,0,-30]; [85,-35,65]; [55,85,-15]; [85,-55,20]; [60,55,-10]
[55,-45,15]; [45,65,-10]; [75,-55,60]; [35,70,-65]; [80,-20,30]
[85,-45,15]; [30,45,-10]; [85,-35,65]; [30,35,-60]; [85,-75,80]
[85,-15,0]; [30,45,35]; [85,-50,0]; [55,65,25]; [75,-50,60]
[70,0,-45]; [85,-80,80]; [30,55,-25]; [70,-40,60]; [45,60,-70]

[85,-15,0]; [30,25,-5]; [85,-55,25]; [40,60,-20]; [85,-50,70]; [45,80,-80]; [55,-35,55]; [65,50,-40]
[85,-30,30]; [50,75,-65]; [70,-65,60]; [45,45,-30]; [85,-35,75]; [35,45,-75]; [60,-35,60]; [70,45,-45]
[70,0,-45]; [80,-55,75]; [45,80,-75]; [65,-25,50]; [30,50,-45]; [55,-55,50]; [55,60,-30]; [80,-55,30]
[85,-45,30]; [50,65,-15]; [85,-70,80]; [30,65,-70]; [85,-40,80]; [30,35,-15]; [60,-40,40]; [55,70,-65]
[85,-30,30]; [50,70,-25]; [65,-60,55]; [50,85,-70]; [60,-25,40]; [35,45,-70]; [85,0,85]; [30,10,-20]

Slider settings:: PD:1.0 ND:0.0 NU:0.0 PP:1.0

[55,-30,0]; [85,-55,10]; [30,25,-45]
[70,-30,15]; [30,-10,5]; [85,-40,60]
[85,-30,-15]; [35,-15,-15]; [75,-45,65]
[85,-30,-15]; [30,-30,25]; [75,-55,40]
[85,-30,-15]; [30,15,-35]; [80,-40,65]

[40,-15,0]; [85,-25,15]; [30,25,-35]; [80,-35,65]; [45,55,-65]
[70,0,-45]; [35,10,-45]; [80,-25,20]; [30,40,-70]; [80,-60,50]
[85,-45,45]; [35,50,-40]; [75,-20,-20]; [30,-5,-25]; [80,-50,0]
[85,-30,0]; [30,15,-35]; [75,-60,45]; [30,50,-25]; [70,10,-35]
[85,-60,45]; [45,65,-40]; [75,-15,-35]; [30,0,-20]; [85,-30,20]

[55,-15,-30]; [85,-35,10]; [30,25,-40]; [85,-45,75]; [40,60,-65]; [55,-35,55]; [80,40,-25]; [30,-25,20]
[55,-45,15]; [85,-10,-20]; [30,40,-70]; [85,-35,60]; [35,-15,-15]; [85,-70,35]; [30,50,-20]; [85,-40,5]
[55,-45,30]; [70,55,-35]; [85,-55,50]; [30,35,-25]; [80,-25,45]; [30,60,-70]; [75,-15,-25]; [30,-20,-5]
[85,-30,0]; [30,25,-55]; [75,-60,50]; [35,60,-35]; [70,0,-35]; [30,-5,-25]; [80,35,-20]; [45,-40,25]
[55,-30,0]; [85,-15,-10]; [30,-25,5]; [85,-55,5]; [30,55,-15]; [70,-50,60]; [40,65,-60]; [85,-80,75]

Slider settings:: PD:1.0 ND:0.0 NU:0.0 PP:0.5

[85,-15,0]; [30,30,-55]; [80,-45,65
[85,-30,-15]; [30,45,-35]; [85,-40,75
[85,-15,-15]; [30,10,-20]; [75,-30,50
[85,-45,60]; [35,60,0]; [85,-35,5
[85,-15,-15]; [30,20,-40]; [85,-50,-5

[70,0,-45]; [30,-25,15]; [85,-30,20]; [30,40,-65]; [70,-40,70
[40,-15,0]; [80,-70,50]; [55,65,-25]; [85,-35,30]; [30,35,-25
[40,0,-30]; [75,-55,45]; [55,70,-35]; [40,-40,30]; [75,40,-20
[70,0,-45]; [30,50,-5]; [75,-60,50]; [55,70,-50]; [80,-30,50
[85,-15,0]; [30,30,5]; [85,-70,40]; [50,75,0]; [85,-45,-10

[85,-15,-15]; [30,-35,25]; [55,85,0]; [75,-55,50]; [50,70,-50]; [80,-25,50]; [30,40,-65]; [80,-50,5
[40,0,-30]; [80,-45,70]; [35,65,-90]; [85,-50,0]; [30,45,-35]; [80-75,60]; [60,80-60]; [45,-35,30
[55,-15,-30]; [75,-55,60]; [30,55,-10]; [80,-50,0]; [55,70,50]; [50,-40,35]; [55,60,-35]; [75,-25,50
[85,-60,30]; [40,60,-15]; [60,-50,45]; [65,45,-40]; [85,-55,75]; [35,70,-90]; [75,-20,40]; [40,35,-65
[85,-45,15]; [45,60,-15]; [80,-50,75]; [35,50,-60]; [85,-15,30]; [40,80,-85]; [50,-45,50]; [70,60,-30

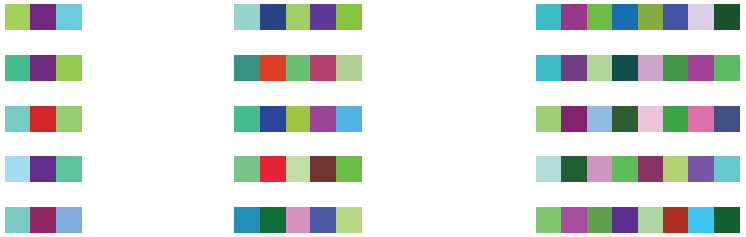
Slider settings:: PD:1.0 ND:1.0 NU:0.0 PP:0.0

[85,-45,15]; [40,35,15]; [75,-40,60]
[85,-45,15]; [45,65,50]; [70,-50,60]
[55,-15,-15]; [30,50,-20]; [85,-65,30]
[85,-60,45]; [30,25,10]; [70,-25,-25]
[85,-30,0]; [45,60,10]; [80,-35,65]

[55,-30,0]; [60,50,0]; [85,-55,15]; [35,35,5]; [80,-55,65]
[40,-15,-15]; [55,60,25]; [65,-45,15]; [30,45,30]; [80,-50,60]
[55,-45,30]; [60,55,0]; [85,-30,60]; [35,55,10]; [85,-40,0]
[85,-30,-15]; [40,55,30]; [60,-50,35]; [65,45,5]; [85,-70,45]
[85,-30,-15]; [40,40,0]; [75,-60,30]; [55,70,30]; [85,-65,80]

[40,0,-30]; [85,-70,55]; [70,50,0]; [50,-35,20]; [45,70,20]; [80,-30,-15]; [35,40,25]; [80,-40,40]
[70,-60,30]; [50,65,-30]; [85,-45,75]; [30,30,-40]; [75,-20,30]; [45,75,-80]; [80,-75,75]; [75,45,-35]
[85,-30,15]; [30,25,5]; [75,-50,65]; [35,60,0]; [80,-30,-30]; [60,45,10]; [55,-45,20]; [45,65,55]
[85,-45,15]; [30,40,-35]; [60,-60,60]; [55,75,-25]; [85,-50,65]; [45,70,-70]; [60,-30,35]; [70,35,-15]
[85,-60,30]; [60,45,-10]; [85,-80,80]; [35,50,-50]; [50,-40,50]; [65,80,-45]; [75,-30,60]; [40,80,-95]

Slider settings:: PD:1.0 ND:1.0 NU:0.0 PP:1.0



[85,-45,60]; [30,45,-35]; [80,-30,-15]
[70,-45,15]; [30,45,-35]; [80,-45,60]
[85,-45,0]; [45,65,55]; [80,-40,45]
[85,-15,-15]; [30,40,-40]; [80,-55,15]
[85,-45,0]; [35,50,-10]; [70,0,-35]

[85,-30,0]; [30,15,-40]; [80,-35,50]; [30,60,-65]; [85,-65,80]
[55,-30,0]; [50,60,60]; [80,-60,40]; [45,50,0]; [80,-20,25]
[70,-45,15]; [30,35,-65]; [75,-30,60]; [45,65,-45]; [70,-10,-45]
[85,-60,30]; [50,70,40]; [85,-20,25]; [30,25,15]; [85,-80,80]
[55,-15,-30]; [40,-45,40]; [70,45,-25]; [40,35,-65]; [85,-30,40]

[70,-30,-15]; [40,50,-25]; [75,-60,65]; [45,0,-40]; [65,-30,60]; [35,60,-90]; [85,10,-15]; [30,-25,15]
[70,-30,-15]; [35,35,-30]; [85,-30,30]; [30,-20,-5]; [75,40,-35]; [55,-45,45]; [45,65,-40]; [80,-70,45]
[85,-45,45]; [30,55,-25]; [75,-5,-25]; [35,-25,20]; [85,30,-20]; [60,-50,40]; [65,65,-25]; [35,10,-30]
[85,-15,0]; [35,-30,20]; [70,40,-25]; [80,-70,50]; [35,40,-10]; [80,-25,45]; [45,50,-60]; [85,-45,-10]
[85,-60,45]; [50,70,-45]; [60,-35,35]; [30,40,-45]; [85,-30,25]; [40,50,40]; [75,-25,-30]; [35,-35,20]

Slider settings:: PD:1.0 ND:1.0 NU:0.0 PP:0.5

[85,-60,45]; [40,50,-15]; [80,-35,0]
[70,0,-45]; [80,-50,50]; [35,65,-45]
[85,-30,30]; [30,30,20]; [80,-60,20]
[85,-60,30]; [45,65,40]; [60,-50,45]
[85,-30,0]; [30,50,35]; [75,-65,45]

[55,-15,-15]; [80,-30,40]; [40,45,-55]; [85,-65,80]; [30,0,-30]
[85,-45,45]; [45,65,-55]; [55,-35,50]; [30,30,-25]; [80,-75,65]
[85,-15,-15]; [50,75,40]; [85,-45,0]; [40,45,30]; [85,-70,45]
[70,-15,-30]; [50,65,-10]; [60,-55,60]; [40,60,-60]; [85,-65,75]
[40,-15,-15]; [75,25,65]; [30,50,5]; [75,-25,0]; [55,75,50]

[85,-60,15]; [55,65,45]; [80,-20,-30]; [35,55,-50]; [85,-25,35]; [30,25,30]; [85,10,0]; [55,-40,45]
[85,-30,0]; [30,25,-15]; [75,-60,65]; [30,30,-60]; [80,-30,50]; [40,70,-85]; [55,-25,50]; [55,80,-30]
[70,-45,15]; [55,70,50]; [75,-60,70]; [35,55,-60]; [75,-30,60]; [65,55,-50]; [40,-15,10]; [80,15,10]
[55,-15,-30]; [85,-60,50]; [35,55,-5]; [85,-20,30]; [30,15,-30]; [80,15,30]; [30,50,-75]; [60,-20,40]
[55,-15,-15]; [40,40,40]; [70,-55,65]; [65,60,-25]; [30,-10,-20]; [85,25,-20]; [35,-40,35]; [50,75,45]

Slider settings:: PD:1.0 ND:0.5 NU:0.0 PP:0.0

[85,-30,0]; [30,25,30]; [80,-45,50]
[85,-30,30]; [30,40,-65]; [85,-40,75]
[85,-30,15]; [55,85,0]; [85,-65,35]
[85,-30,0]; [30,45,30]; [85,-55,45]
[40,0,-30]; [80,-30,45]; [30,50,-75]

[85,-30,0]; [30,40,25]; [80,-65,40]; [45,70,25]; [65,-30,55]
[85,-30,30]; [55,65,-10]; [80,-60,45]; [30,30,-20]; [70,-25,-30]
[85,-45,30]; [30,50,-20]; [85,-40,80]; [55,65,-35]; [85,-75,60]
[85,-45,15]; [30,50,0]; [80,-50,65]; [55,85,0]; [55,-50,55]
[85,-30,30]; [35,55,-20]; [60,-50,60]; [40,60,-65]; [85,-40,80]

[85,-45,15]; [50,80,-15]; [75,-60,70]; [45,45,-50]; [50,-40,45]; [55,80,-70]; [75,-30,60]; [30,70,-95]
[85,-30,30]; [45,50,-25]; [75,-60,70]; [45,70,-70]; [50,-25,35]; [75,55,-35]; [80,-25,75]; [50,40,-75]
[85,-60,45]; [60,75,-35]; [60,-40,40]; [35,45,-25]; [85,-25,45]; [40,70,-80]; [55,25,50]; [75,-5,-30]
[70,0,-45]; [30,45,-5]; [65,-50,40]; [50,80,-15]; [85,-80,80]; [45,75,-60]; [40,-40,35]; [55,40,-20]
[40,0,-30]; [80,-25,45]; [40,65,-25]; [65,-65,65]; [65,55,-35]; [50,-35,35]; [60,95,-60]; [40,30,45]

Slider settings:: PD:1.0 ND:0.5 NU:0.0 PP:1.0

[85,-15,-15]; [30,60,-55]; [85,-50,5]
[85,-45,45]; [35,60,-55]; [80,-40,-10]
[85,-45,45]; [30,45,-40]; [85,-15,-20]
[85,-30,0]; [35,45,35]; [85,-60,20]
[70,-30,0]; [30,50,35]; [80,-35,45]

[85,-30,30]; [30,45,-75]; [85,-45,75]; [30,5,-20]; [85,-35,-15]
[85,-30,15]; [30,50,-30]; [70,-25,-30]; [30,20,-20]; [75,-55,60]
[70,-30,15]; [45,65,-30]; [70,-60,60]; [35,15,-45]; [80,35,-25]
[85,-60,15]; [30,55,-50]; [80,-30,-15]; [30,20,-40]; [80,-25,35]
[85,-45,60]; [30,55,-15]; [80,-10,-30]; [30,-5,-20]; [80,-40,10]

[40,-15,0]; [85,-35,-20]; [45,65,-10]; [75,-70,65]; [35,70,-65]; [80,-35,50]; [30,30,-55]; [55,-25,50]
[40,0,-30]; [85,-60,75]; [30,55,-55]; [75,-60,30]; [60,70,-25]; [75,-35,-20]; [35,30,-10]; [85,0,-20]
[85,-60,30]; [45,65,-15]; [75,-35,-20]; [30,20,-90]; [80,-30,50]; [30,55,-75]; [80,35,-25]; [55,-45,50]
[70,0,-45]; [35,60,-50]; [70,-50,45]; [30,30,-45]; [80,-35,-20]; [40,-40,25]; [60,75,-20]; [40,0,-40]
[70,-30,15]; [30,55,-5]; [85,-60,70]; [40,80,-85]; [85,-30,60]; [40,45,-55]; [80,-25,-30]; [30,-30,15]

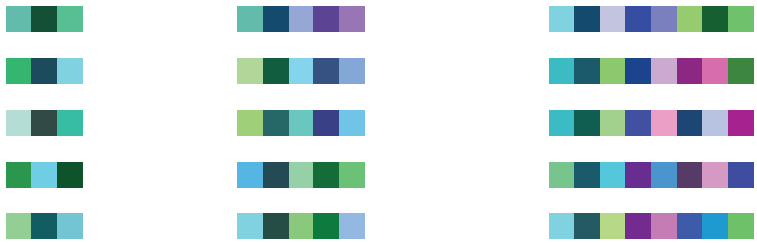
Slider settings:: PD:1.0 ND:0.5 NU:0.0 PP:0.5

[85,-15,-15]; [35,50,45]; [70,-55,20]
[85,-45,15]; [40,55,45]; [85,-55,60]
[85,-15,-15]; [30,40,-45]; [80,-20,40]
[70,-30,0]; [30,40,40]; [85,-80,80]
[85,-45,60]; [35,50,0]; [85,-25,0]

[55,-15,-15]; [85,-35,60]; [30,45,-50]; [60,-50,60]; [70,60,-35]
[55,-15,-15]; [80,-65,50]; [70,60,-45]; [75,-35,35]; [35,50,-45]
[55,-45,30]; [50,70,-45]; [80,-35,75]; [30,15,-40]; [85,-50,30]
[70,0,-45]; [30,-35,25]; [55,70,-35]; [70,-35,50]; [30,40,-70]
[85,-60,15]; [35,35,5]; [80,-45,75]; [55,80,10]; [75,-30,25]

[85,-30,15]; [30,45,-5]; [85,-75,50]; [50,75,-45]; [55,-40,55]; [30,45,-50]; [85,-35,75]; [30,15,-30]
[55,-15,-30]; [35,45,-20]; [75,-50,70]; [45,85,-85]; [80,-20,25]; [30,-5,-25]; [75,40,-15]; [50,-45,45]
[55,-15,-30]; [30,40,-5]; [80,-15,-25]; [40,70,-35]; [85,-75,60]; [50,70,40]; [60,-45,40]; [65,50,-15]
[70,0,-45]; [30,-35,30]; [70,45,-30]; [80,-70,70]; [55,80,-65]; [85,-40,75]; [30,60,-50]; [85,-40,10]
[70,-30,15]; [35,60,5]; [85,-60,55]; [50,60,-45]; [30,-35,30]; [75,35,-25]; [60,-30,60]; [30,15,-45]

Slider settings:: PD:0.5 ND:0.0 NU:0.0 PP:1.0

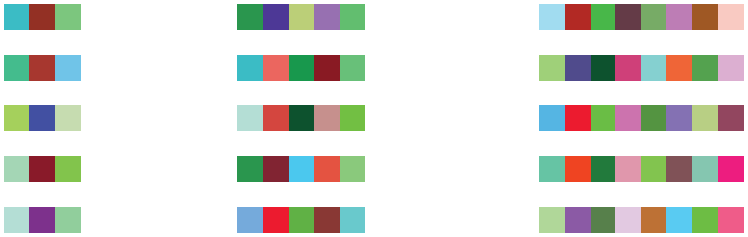


[70,-30,0]; [30,-25,10]; [80,-60,20]
[70,-60,30]; [30,-10,-15]; [85,-30,-15]
[85,-15,0]; [30,-10,0]; [75,-50,5]
[55,-45,30]; [85,-35,-20]; [30,-35,25]
[85,-45,30]; [35,-20,-10]; [75,-20,-15]

[70,-30,0]; [30,-5,-25]; [70,10,-40]; [35,30,-40]; [60,50,-55]
[85,-30,30]; [35,-30,10]; [85,-25,-20]; [35,5,-30]; [70,5,-45]
[85,-45,45]; [40,-20,-5]; [85,-50,0]; [30,20,-40]; [75,-15,-25]
[70,-15,-30]; [30,-10,-10]; [85,-40,20]; [40,-40,30]; [85,-70,40]
[85,-30,-15]; [30,-15,0]; [85,-55,40]; [45,-45,30]; [75,0,-35]

[85,-30,-15]; [30,-5,-25]; [80,5,-15]; [35,25,-55]; [60,30,-60]; [80,-40,45]; [35,-35,25]; [85,-70,45]
[70,-30,-15]; [35,-15,-15]; [80,-45,45]; [30,15,-45]; [75,25,-25]; [35,55,-30]; [65,70,-30]; [50,-40,40]
[70,-30,-15]; [35,-25,0]; [85,-40,35]; [35,40,-70]; [75,40,-15]; [30,5,-30]; [80,5,-25]; [40,70,-35]
[85,-60,30]; [35,-15,-15]; [80,-35,-20]; [30,50,-50]; [60,-5,-35]; [30,20,-20]; [70,30,-15]; [30,70,-95]
[85,-30,-15]; [35,-15,-10]; [85,-30,40]; [30,60,-50]; [65,55,-35]; [40,30,-65]; [60,-15,-35]; [80,-60,45]

Slider settings:: PD:0.5 ND:1.0 NU:0.0 PP:0.0



[70,-30,-15]; [35,40,30]; [80,-50,35]
[70,-45,15]; [40,45,30]; [75,-15,-25]
[85,-45,60]; [35,40,-70]; [85,-15,20]
[85,-30,15]; [30,45,20]; [80,-55,60]
[85,-15,0]; [35,45,-35]; [85,-45,25]

[55,-45,30]; [30,40,-55]; [80,-20,40]; [55,35,-40]; [80,-65,40]
[70,-30,-15]; [60,50,30]; [55,-50,30]; [30,45,25]; [80,-60,35]
[85,-15,0]; [50,55,35]; [30,-30,15]; [65,20,10]; [70,-45,55]
[55,-45,30]; [30,40,15]; [80,-30,-30]; [55,55,40]; [85,-55,40]
[70,0,-45]; [50,75,45]; [65,-45,50]; [35,35,20]; [80,-35,-10]

[85,-15,-15]; [40,55,45]; [70,-65,60]; [30,20,0]; [65,-30,30]; [65,55,-40]; [45,25,50]; [85,15,10]
[85,-45,45]; [35,20,-35]; [30,-30,15]; [50,60,0]; [85,-35,-5]; [60,50,50]; [60,-40,35]; [80,40,-30]
[70,-15,-30]; [50,75,45]; [85,-80,70]; [65,65,-35]; [55,-40,55]; [55,35,-50]; [80,-20,35]; [40,35,0]
[85,-60,15]; [55,65,65]; [45,-40,30]; [70,30,0]; [85,-65,60]; [40,20,5]; [75,-25,5]; [55,80,5]
[85,-30,30]; [50,55,-55]; [50,-25,25]; [85,20,-20]; [55,25,45]; [80,-25,-30]; [80,-70,75]; [60,60,5]

Slider settings:: PD:0.5 ND:1.0 NU:0.0 PP:1.0

[85,-15,-15]; [35,-20,15]; [85,-50,5]
[70,-30,0]; [30,-35,30]; [65,-5,-45]
[85,-15,-15]; [40,-35,20]; [85,-45,-5]
[85,-60,15]; [45,50,-50]; [85,-15,-5]
[40,-15,0]; [85,-25,50]; [55,0,-50]

[85,-45,15]; [30,15,-25]; [85,-50,60]; [35,60,-55]; [70,-15,-25]
[85,-45,30]; [35,35,-15]; [85,-35,75]; [30,-5,-15]; [75,-15,-25]
[85,-30,-15]; [40,-40,30]; [60,0,-45]; [65,-50,50]; [45,60,-45]
[85,-30,15]; [30,25,-50]; [70,60,-30]; [35,-20,0]; [85,0,-20]
[85,-30,15]; [40,30,-35]; [60,-40,50]; [40,65,-45]; [70,-10,-45]

[85,-15,-15]; [40,-40,25]; [85,-45,0]; [30,0,-20]; [70,65,-45]; [55,-15,-20]; [35,55,-15]; [85,-15,30]
[40,0,-30]; [75,-40,45]; [45,60,-45]; [80,-25,-15]; [30,-30,15]; [80,30,-30]; [35,30,-60]; [70,70,-45]
[85,-30,-15]; [35,-35,20]; [70,10,-40]; [35,45,-40]; [75,-45,50]; [45,5,-40]; [65,50,-40]; [60,-25,-10]
[85,-45,15]; [30,25,-20]; [85,-35,75]; [30,60,-50]; [65,-15,-30]; [45,-45,30]; [70,35,-35]; [40,30,-70]
[55,-45,30]; [75,-20,-25]; [30,30,-35]; [85,-55,75]; [45,60,-45]; [60,15,-60]; [30,-30,25]; [85,-55,15]

Slider settings:: PD:0.5 ND:1.0 NU:0.0 PP:0.5

[85,-15,-15]; [30,45,0]; [85,-50,15]
[85,-45,45]; [30,45,20]; [80,-40,-10]
[85,-45,60]; [40,60,30]; [80,-40,5]
[40,0,-30]; [75,-25,30]; [35,40,-60]
[85,-15,0]; [40,30,25]; [85,-50,-5]

[85,-15,0]; [30,25,-35]; [85,-65,65]; [60,50,-25]; [55,-40,40]
[70,-45,15]; [35,20,20]; [65,-35,60]; [50,60,35]; [85,-15,-15]
[85,-45,60]; [50,65,10]; [65,-35,0]; [55,35,55]; [80,-75,75]
[85,-60,45]; [40,10,-50]; [75,-20,35]; [50,50,-35]; [75,-35,-15]
[70,0,-45]; [30,55,-30]; [80,-60,40]; [40,15,-20]; [70,-30,40]

[85,-45,30]; [30,55,-65]; [85,-60,65]; [40,55,0]; [80,-10,10]; [55,70,-50]; [55,-45,40]; [40,20,-35]
[85,-60,30]; [40,35,-30]; [85,-75,80]; [40,-5,-25]; [85,-40,80]; [70,50,-45]; [55,-20,20]; [80,-25,-30]
[85,-15,-15]; [35,45,25]; [65,-60,60]; [40,40,-50]; [70,-45,10]; [50,75,10]; [30,-15,5]; [70,15,-45]
[85,-15,0]; [35,30,20]; [65,-40,50]; [45,55,-50]; [85,-75,70]; [30,20,-35]; [85,-50,5]; [60,60,30]
[85,-60,45]; [30,65,-75]; [60,-40,60]; [60,0,-45]; [30,55,-20]; [85,-15,-5]; [35,-25,15]; [70,55,-40]

Slider settings:: PD:0.5 ND:0.5 NU:0.0 PP:1.0

[85,-15,-15]; [40,-30,20]; [85,-45,-10]
[85,-15,-15]; [30,-15,0]; [85,-45,-10]
[55,-45,30]; [85,-25,-10]; [30,-15,0]
[70,-30,-15]; [35,-35,20]; [85,-45,30]
[85,-60,15]; [30,45,-55]; [85,-20,-5]

[85,-60,45]; [40,10,-40]; [70,-25,35]; [35,55,-35]; [80,-5,-30]
[55,-15,-15]; [85,-60,45]; [35,50,-25]; [85,-50,0]; [30,0,-30]
[85,-45,30]; [30,25,-35]; [75,10,-35]; [30,60,-50]; [85,-20,-20]
[70,-15,-30]; [35,45,-20]; [70,-45,5]; [30,-5,-25]; [80,35,-25]
[85,-45,60]; [35,35,-65]; [60,-30,60]; [70,-15,-35]; [30,-30,20]

[55,-15,-30]; [85,-50,70]; [30,-10,-20]; [85,-10,-15]; [35,-40,35]; [85,-40,10]; [30,40,-45]; [60,-45,20]
[55,-15,-15]; [80,-55,50]; [45,70,-45]; [55,-55,45]; [80,-35,-15]; [30,-20,15]; [85,-25,30]; [35,25,-55]
[40,-15,0]; [85,-55,40]; [30,15,-45]; [75,50,-35]; [30,45,-70]; [80,15,-30]; [45,-45,35]; [75,-15,-25]
[85,-30,30]; [40,-15,-15]; [80,-60,55]; [35,50,10]; [70,-25,-25]; [40,-40,30]; [80,20,30]; [35,20,15]
[85,-45,60]; [30,40,-50]; [80,5,-20]; [40,-25,-5]; [85,-35,-10]; [35,5,-30]; [70,50,-45]; [55,-35,55]

Measure	Colors	R^2	$F(3, 13196)$	p	β_{PD}	t_{PD}	p_{PD}	β_{ND}	t_{ND}	p_{ND}	β_{NU}	t_{NU}	p_{NU}	β_{PP}	t_{PP}	p_{PP}	RI_{PD}	RI_{ND}	RI_{NU}	RI_{PP}
PD	3	0.492	4262.806	*	NA	NA	NA	0.16	36.358	*	-0.083	-24.822	*	-0.165	-23.019	*	NA	0.248	0.052	0.191
ND	3	0.738	12411.646	*	0.571	36.358	*	NA	NA	NA	-0.215	-34.622	*	-1.138	-117.653	*	0.233	NA	0.043	0.461
NU	3	0.186	1007.676	*	-0.537	-24.822	*	-0.388	-34.622	*	NA	NA	NA	-0.646	-36.408	*	0.065	0.076	NA	0.043
PP	3	0.701	10310.915	*	-0.234	-23.019	*	-0.45	-117.653	*	-0.141	-36.408	*	NA	NA	NA	0.193	0.489	0.018	NA
PD	5	0.343	2295.15	*	NA	NA	NA	0.087	26.478	*	-0.038	-16.259	*	-0.206	-38.652	*	NA	0.15	0.024	0.167
ND	5	0.508	4537.629	*	0.579	26.478	*	NA	NA	NA	-0.263	-47.172	*	-0.884	-72.06	*	0.143	NA	0.086	0.277
NU	5	0.2	1096.931	*	-0.521	-16.259	*	-0.549	-47.172	*	NA	NA	NA	-0.808	-41.001	*	0.028	0.117	NA	0.053
PP	5	0.486	4155.721	*	-0.494	-38.652	*	-0.319	-72.06	*	-0.14	-41.001	*	NA	NA	NA	0.162	0.286	0.036	NA
PD	8	0.328	2148.682	*	NA	NA	NA	0.042	17.753	*	-0.058	-41.791	*	-0.14	-40.325	*	NA	0.1	0.108	0.119
ND	8	0.387	2775.101	*	0.561	17.753	*	NA	NA	NA	-0.211	-41.174	*	-0.701	-58.12	*	0.099	NA	0.093	0.193
NU	8	0.274	1662.23	*	-2.011	-41.791	*	-0.54	-41.174	*	NA	NA	NA	-0.866	-42.72	*	0.115	0.105	NA	0.052
PP	8	0.369	2577.387	*	-0.785	-40.325	*	-0.291	-58.12	*	-0.14	-42.72	*	NA	NA	NA	0.121	0.201	0.046	NA

S.Table 4. Linear regression tables for the 12 Palette Verification regressions that predicted palette scores in terms of the other three. Relative importance is calculated with `lmg` in the "relaimpo" R package [1]. $*: p < 0.001$

Measure	Colors	R^2	$F(3, 16)$	p	β_{PD}	t_{PD}	p_{PD}	β_{ND}	t_{ND}	p_{ND}	β_{PP}	t_{PP}	p_{PP}	RI_{PD}	RI_{ND}	RI_{PP}
RT	3	0.815	23.442	< 0.001	-137.624	-4.485	< 0.001	-165.412	-5.39	< 0.001	17.627	0.574	0.574	0.211	0.357	0.246
RT	5	0.682	11.447	< 0.001	-81.807	-2.84	0.012	-126.571	-4.394	< 0.001	-2.646	-0.092	0.928	0.113	0.411	0.158
RT	8	0.192	1.267	0.319	-5.719	-0.411	0.687	-22.447	-1.613	0.126	-18.198	-1.308	0.209	0.029	0.108	0.054
Error	3	0.712	13.186	< 0.001	-0.488	-2.251	0.039	-0.85	-3.921	0.001	0.227	1.049	0.31	0.09	0.375	0.248
Error	5	0.692	11.964	< 0.001	-0.486	-2.001	0.063	-0.934	-3.85	0.001	0.224	0.924	0.369	0.073	0.39	0.228
Error	8	0.537	6.192	0.005	-0.267	-1.157	0.264	-0.568	-2.466	0.025	0.241	1.048	0.31	0.049	0.277	0.211
Pref. Rating	3	0.868	35.089	< 0.001	-6.754	-1.068	0.301	-20.665	-3.268	0.005	32.271	5.103	< 0.001	0.067	0.275	0.526
Pref. Rating	5	0.581	7.396	0.003	5.527	2.208	0.042	-4.013	-1.603	0.128	2.959	1.182	0.254	0.235	0.282	0.065
Pref. Rating	8	0.495	5.228	0.01	-0.876	-0.103	0.919	-15.615	-1.837	0.085	12.843	1.511	0.15	0.021	0.254	0.22

S.Table 5. Linear regression tables for Experiment 1 that predicted participants' Response Time, Error, and Preference Rating as a function of Perceptual Distance (PD), Name Difference (ND), and Pair Preference (PP) slider settings. Relative importance is calculated with `lmg` in the "relaimpo" R package [1].

REFERENCES

[1] U. Groemping. Relative importance for linear regression in r: The package relaimpo. *Journal of Statistical Software*, 17(1):1–27, 2006.

[2] J. Heer and M. Stone. Color naming models for color selection, image editing and palette design. In *ACM Human Factors in Computing Systems (CHI)*, pages 1007–1016, New York, NY, USA, 2012. ACM.

[3] X.-L. Meng, R. Rosenthal, and D. B. Rubin. Comparing correlated correlation coefficients. *Psychological bulletin*, 111(1):172, 1992.

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