Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical

Angela M. Brown
College of Optometry, The Ohio State University, Columbus, OH, USA

Delwin T. Lindsey
Department of Psychology, The Ohio State University, Mansfield, OH, USA

Kevin M. Guckes
College of Optometry, The Ohio State University, Columbus, OH, USA

The relation between colors and their names is a classic case study for investigating the Sapir–Whorf hypothesis that categorical perception is imposed on perception by language. Here, we investigate the Sapir–Whorf prediction that visual search for a green target presented among blue distractors (or vice versa) should be faster than search for a green target presented among distractors of a different color of green (or for a blue target among different blue distractors). A. L. Gilbert, T. Regier, P. Kay, and R. B. Ivry (2006) reported that this Sapir–Whorf effect is restricted to the right visual field (RVF), because the major brain language centers are in the left cerebral hemisphere. We found no categorical effect at the Green–Blue color boundary and no categorical effect restricted to the RVF. Scaling of perceived color differences by Maximum Likelihood Difference Scaling (MLDS) also showed no categorical effect, including no effect specific to the RVF. Two models fit the data: a color difference model based on MLDS and a standard opponent-colors model of color discrimination based on the spectral sensitivities of the cones. Neither of these models nor any of our data suggested categorical perception of colors at the Green–Blue boundary, in either visual field.

Keywords: color vision, categorization, search, color appearance/constancy, eye movements


Introduction

The Sapir–Whorf hypothesis has been influential in the fields of psychology, philosophy, anthropology, and linguistics. According to the Sapir–Whorf hypothesis, our perception of stimuli depends on the names we give to them. Following the classic definition of categorical perception (reviewed in Harnad, 1987), stimuli within categories are perceptually similar to one another and are given the same name, whereas stimuli from different categories look different and are given different names. According to the Sapir–Whorf hypothesis, it is the distinctive names that are responsible for the categorical perception of stimuli.

Color names are a classic topic of research within the tradition of the Sapir–Whorf hypothesis. Almost all world languages have at least some color names in their lexicons, yet there is important variation around the world in how many color terms are used (Kay, Berlin, Maffi, Merrifield, & Cook, 2009). Furthermore, almost all people have color vision, which can be easily measured in the laboratory and in the field. Finally, the spectral composition of light is continuously variable, so any categorical perception of colors that might exist must be due to the perception of the observer rather than to the stimuli themselves. Colors are easy to specify, measure, and produce, which allows the relations among the physical properties of colors, the perception of colors, and the naming of colors to rest on a quantitative basis.

Many investigators have attempted to link the color categories defined by color names to other aspects of color-based visual performance such as judged similarity (Kay & Kempton, 1984) and visual search (Daoutis, Franklin, Riddett, Clifford, & Davies, 2006; Lindsey, Brown, Reijnjen, Rich, Kuzmova, & Wolfe, 2010) and to visual physiology more generally (Lindsey & Brown, 2002; Ratliff, 1976). Particularly, a recent paper by Gilbert, Regier, Kay, and Ivry (2006; henceforth “Gilbert et al.”) claims that visual search reaction time for colors is about 0.024 s faster across the Green–Blue color category boundary than in the case of stimuli that are within the Green category or within the Blue category but only for stimuli presented in the right visual field (RVF). Gilbert et al. found no category effect in the left visual field (LVF). This result suggested in a general way that the language centers in the left cerebral hemisphere are important. Here, we attempt to replicate and extend that result.
Overview of the experiments

Experiments I–IV were visual search reaction time (RT) studies in which the observer viewed a circle of 12 colored stimuli (Figures 1a and 1b) and identified the odd one out. Classically, if stimuli are perceived categorically, targets are found faster when they are in a different category from the distractors than when they are in the same category as the distractors. We did not find evidence of categorical perception that is restricted to the right visual field. In Experiment I, the target and distractor stimuli were similar to those used in Gilbert et al. Our stimuli replicated those used in Kay and Kempton (1984), being at constant value and chroma, but varying in hue, within the Munsell color ordering system. As in Gilbert et al., the visual search data were button-press reaction times. In Experiment II, the stimuli were at constant distance from white in CIELAB 1976, a color space based on the discriminability of colors (Schanda, 2007), and the search data were saccadic RTs as the observer looked at the target. In Experiment III, the stimuli were a subset of those used in Experiment II, but the method was button-press RT. In Experiment IV, the stimuli were also at constant eccentricity in CIELAB, while two technical aspects of the experiment were manipulated: (1) either the surrounding field was dark, as in our Experiments I–III, or light, as it was in Gilbert et al., and (2) saccadic RTs were collected as the observer looked at the target or looked at a “button” on the right- or left-hand side of the display, according to whether the target was on the right or left side of the fixation point. We compared the color at the RT minimum from Experiments I–IV to each observer’s Green–Blue boundary and found that the fastest color was not related to the Green–Blue boundary.

In Experiment V, we addressed the question of whether these experiments provide evidence of categorical perception. We collected scaling data using the Maximum Likelihood Difference Scaling (MLDS) method of Maloney and Yang (2003) and used the scaled color differences to predict the shape of the RT data sets. Those data agreed well, suggesting that the RT data were related to the perceived differences between the colors: RT was shorter when the colors looked more different, and RT was longer when the colors looked more similar. Finally, we modeled the RT data using a standard color-opponent model (Lindsey et al., 2010). We found no particularly large discrepancy between this standard model and the data in the neighborhood of the Green–Blue boundary, such as that predicted by Gilbert et al. and others. Taken as a whole, the results and analyses suggested that the overall shape of the RT data sets was controlled entirely by visual signals that arise in the cones and are combined in a color-opponent fashion in the earliest stages of visual processing.

General methods for Experiments I–IV

These studies were carried out under the tenets of the Declaration of Helsinki and were approved by the Biomedical Institutional Review Board of the Ohio State University. All subjects participated after informed written consent was obtained.

The basic stimulus configuration was a ring of 12 color samples, eleven of which were distractors of one color and one of which was a target of an “odd” color, which was either greener or bluer than the eleven distractors (Figures 1a and 1b and Appendix A; calibrations performed with a Pritchard PR-670 SpectraScan spectrophotometer).

The ring of colored samples (disks in Experiments I and III; squares in Experiments II and IV) appeared around the
fixation point and remained present until the observer responded. The observer's task was to determine which of the 12 stimuli was the “odd one,” i.e., the target. In Experiments I and III, the observer responded promptly by pressing a button to indicate whether the “odd one” was to the right or left of the fixation point. Color names were never used in the instructions for the RT experiment, and observers generally did not know that color terms or color categories were of interest to the experimenters, except for coauthors AMB and KMG, whose data are clearly marked in every figure where they are reported. The visual-search button-press reaction time (RT) was the time elapsed between the appearance of the ring of stimuli and the button press, on correct trials only. In Experiments II and IV, the observer looked directly at the target, and the saccadic RT was the time between the onset of the stimulus and the arrival of the line of sight at the correct target stimulus after a single saccadic eye movement. The other details of the stimuli are in Appendix A. The RT results are expressed in units of seconds, as a function of the color azimuth in CIELAB.

The experimental hypothesis, based on Gilbert et al. and Kay and Kempton (1984), was that reaction times for the between-color-category stimulus combinations should be shorter than for the within-color-category combinations, but that this Whorfian effect should apply only to the right visual field. Inasmuch as this hypothesis concerned whether colors were categorically green or blue, each of our experiments included a measurement of each observer’s Green–Blue boundary, which only occurred after the RT phase of the experiment was completed. The stimulus for this phase of each experiment was the same ring of disks or squares used for the RT phase of the experiment. Eleven of the stimuli were one color and one was a different color, either greener or bluer than the other eleven. The separation in color space between the colors was close to the same as that used in the corresponding RT experiment. On a given method-of-adjustment (MOA) trial, the subject used the computer mouse to adjust the colors so that the Green–Blue boundary was halfway between the two colors. This was true when the colors straddled the Green–Blue boundary and the greener stimulus looked as green as the bluer stimulus looked blue (or vice versa, in the case of a bluer target and greener distractors). While the Method of Constant Stimuli has a better reputation among psychophysicists as a method of estimating threshold, we found a pronounced effect of the stimulus range on the measured Green–Blue boundary in pilot experiments (Appendix B), which would likely have biased our results. This relation between the stimulus range and measured boundary was not evident in our MOA data, so that is the method we chose.

Throughout this paper, colors are specified by their azimuth angles within a constant-luminance plane in CIELAB 1976 Uniform Color Space (Schanda, 2007), relative to the point (1, 0). RT and MOA data are plotted at the average of the target and distractor azimuth values. A preliminary Mixed-Procedure Analysis of Variance (SPSS) revealed that there were no statistically reliable differences between the target-greener and target-bluer stimulus combinations for a given average azimuth value, so the data for those two stimulus combinations were pooled throughout. Each RT data point is plotted at the average of the azimuth values of the targets and distractors.

### Experiment I

Experiment I was designed to replicate the basic result of Gilbert et al. by relating each observer’s Green–Blue boundary to his/her search behavior in a button-press latency experiment. The target and distractor colors were at value 5 and chroma 6 of the Munsell color order system and were separated by 5 Munsell hue steps (chromaticities shown in Appendix A). The first four colors (chroma = 7.5G, 2.5BG, 7.5BG, 2.5B) were from Kay and Kempton (1984), as were the intended colors in Gilbert et al. The fifth color was bluer (chroma = 7.5B). To express the results of Experiment I relative to an interval-scaled numerical axis, we calculated the positions of our colors in the CIELAB 1976 Uniform Color Space, and we express our results in units of color azimuth within an isoluminant plane in that space. Because of the way the Munsell color ordering system was constructed, the separation in units of color azimuth in CIELAB space, and separation in ΔE units in CIELAB space, varied from stimulus pair to stimulus pair (minimum = 10.4, maximum = 14.8). The colors were presented on a CRT video monitor (Appendix A).

Fifteen monolingual native English-speaking adults (ages 20–62, median = 41) served as observers. All were right-handed (7 females) and all had normal color vision (D-15 panel). Two were aware of the purposes of the experiment (coauthors AMB and KMG), and the others were experimentally naive. The observer’s task was to find the target and press the “S” key on the computer keyboard with the left hand if the target was on the left half of the display or the “L” key with the right hand if the stimulus was on the right half of the display. The computer recorded the subject’s RT. Each observer contributed eight iterations through the stimulus set (a total of 960 trials).

### Results

On average, observers chose the correct stimulus on 98.6% of trials. Each observer’s Green-Blue boundary value (Figure 2) is indicated by a black triangle just below
his/her RT data set. The observers’ average Green–Blue boundary was at 183.46° of azimuth in CIELAB color space (standard deviation = 9.13°). This is within the range reported by Gilbert et al. and did not differ significantly for the LVF compared to the RVF ($t_{14} = 0.736$, N.S.; average data in Figure 3a).

According to the classic definition of categorical color perception, if color perception had been categorical in this experiment, the fastest RT should have occurred at the Green–Blue boundary. To determine whether this was the case, we estimated the color azimuth of the fastest color perception, if color perception had been categorical in this experiment. The results from the green target/green distractor (average azimuth = 173.76°) and the results from the blue target/blue distractor (average azimuth = 207.97°) were averaged into a single “within-category” group. The within-category results were compared to the between-category group, consisting of the trials with the green target and blue distractors and the blue target and green distractors (average azimuth = 189.16°). Thus, the ANOVA had factors for the visual field of presentation (L–R), the categories of target and distractors (Within vs. Between (W–B) category groups), and Subjects (SUBJ).

The analysis revealed a highly statistically significant effect of W–B ($F(1,14) = 30.101, p < 0.0005$, Figure 4). This result indicates that the data showed a statistically significant minimum near the color azimuth of 189.16° in CIELAB. There was also a significant overall effect of L–R (stimuli in the LVF were slightly faster: $F(1,14) = 6.054, p = 0.027$), but there was no statistically significant interaction between L–R and W–B ($F(1,14) = 0.008, p = 0.932$) and no interaction between those factors and SUBJ. We repeated the analysis based on different assumptions, by considering the green target/green distractor separately from the blue target/blue distractor and by using each subject’s own Green–Blue boundary instead of the group average. Neither of these variations yielded results materially different from the original. Thus, when we repeated the statistical analysis of Gilbert et al., on data that were collected using their reported stimuli and method, we saw no evidence that between-category target stimuli are found faster in the RVF than in the LVF.

To further examine the statistical power of our negative result, we performed an SPSS Mixed-Procedure Analysis of Variance specifically aimed at determining whether our data excluded the 0.024-s advantage that Gilbert et al. reported for between-category stimulus color combinations, within the RVF only. This experimental hypothesis depended on whether the target was in the same color category as the distractors, but subjects differed in their Green–Blue color boundary locations. Therefore, we subtracted each subject’s color boundary azimuth value from the color azimuths of the stimuli. The experimental hypothesis concerned whether the RT was faster in the RVF than in the LVF, so we subtracted each subject’s LVF RT from his/her RVF RT. Thus, we created a data set that had bluer stimuli with positive “color difference”

(r = 0.256, $p = 0.1779$, one-tailed; Figure 3b). Thus, the minimum RT values of the data set were not strongly related to the observers’ measured Green–Blue boundaries.

## Statistical analyses

The first analysis was designed to determine whether the results agreed with those of Gilbert et al. This was a GLM (SPSS) analysis of the data for the three conditions that were the same for this experiment and Gilbert et al.’s experiment. The results from the green target/green distractor (average azimuth = 173.76°) and the results from the blue target/blue distractor (average azimuth = 207.97°) were averaged into a single “within-category” group. The within-category results were compared to the between-category group, consisting of the trials with the green target and blue distractors and the blue target and green distractors (average azimuth = 189.16°). Thus, the ANOVA had factors for the visual field of presentation (L–R), the categories of target and distractors (Within vs. Between (W–B) category groups), and Subjects (SUBJ).

The analysis revealed a highly statistically significant effect of W–B ($F(1,14) = 30.101, p < 0.0005$, Figure 4). This result indicates that the data showed a statistically significant minimum near the color azimuth of 189.16° in CIELAB. There was also a significant overall effect of L–R (stimuli in the LVF were slightly faster: $F(1,14) = 6.054, p = 0.027$), but there was no statistically significant interaction between L–R and W–B ($F(1,14) = 0.008, p = 0.932$) and no interaction between those factors and SUBJ. We repeated the analysis based on different assumptions, by considering the green target/green distractor separately from the blue target/blue distractor and by using each subject’s own Green–Blue boundary instead of the group average. Neither of these variations yielded results materially different from the original. Thus, when we repeated the statistical analysis of Gilbert et al., on data that were collected using their reported stimuli and method, we saw no evidence that between-category target stimuli are found faster in the RVF than in the LVF.

The experiment was designed to determine whether the results agreed with those of Gilbert et al. This was a GLM (SPSS) analysis of the data for the three conditions that were the same for this experiment and Gilbert et al.’s experiment. The results from the green target/green distractor (average azimuth = 173.76°) and the results from the blue target/blue distractor (average azimuth = 207.97°) were averaged into a single “within-category” group. The within-category results were compared to the between-category group, consisting of the trials with the green target and blue distractors and the blue target and green distractors (average azimuth = 189.16°). Thus, the ANOVA had factors for the visual field of presentation (L–R), the categories of target and distractors (Within vs. Between (W–B) category groups), and Subjects (SUBJ). The analysis revealed a highly statistically significant effect of W–B ($F(1,14) = 30.101, p < 0.0005$, Figure 4). This result indicates that the data showed a statistically significant minimum near the color azimuth of 189.16° in CIELAB. There was also a significant overall effect of L–R (stimuli in the LVF were slightly faster: $F(1,14) = 6.054, p = 0.027$), but there was no statistically significant interaction between L–R and W–B ($F(1,14) = 0.008, p = 0.932$) and no interaction between those factors and SUBJ. We repeated the analysis based on different assumptions, by considering the green target/green distractor separately from the blue target/blue distractor and by using each subject’s own Green–Blue boundary instead of the group average. Neither of these variations yielded results materially different from the original. Thus, when we repeated the statistical analysis of Gilbert et al., on data that were collected using their reported stimuli and method, we saw no evidence that between-category target stimuli are found faster in the RVF than in the LVF.
(x-axis) numbers, greener stimuli with negative color difference numbers, and color combinations that straddled the Green–Blue boundary at or near color difference = 0 (Figure 3c). If the RVF were selectively faster at the Green–Blue boundary, the RVF–LVF “RT difference” would be at its minimum value at the Green–Blue boundary. When the data were presented in this way, the experimental question can be expressed as follows: Was the RT difference function U-shaped, with its minimum located at color difference $x = 0$ and the RT difference $y = -0.24$, the difference reported by Gilbert et al.?

The confidence interval containing the RVF–LVF RT difference at the Green–Blue boundary was 0.026142–0.003446, a marginally statistically significant effect in the

Figure 3. Analyses of the results of Experiments I, II, and III. (a) Average data from Figure 2, ±1 SEM. Black triangle: the average Green–Blue boundary. Dashed line: the colors that were used in the statistical analysis of the local minima. If the RT minimum occurred reliably at the Green–Blue boundary, the two measures would be equal and highly correlated (dashed line). Instead, they were uncorrelated (solid line) and the RT minimum was at a bluer color azimuth than the Green–Blue boundary. (c) RT difference as a function of color difference for Experiment I; see text for description of the axis units. The prediction from Gilbert et al. is that the minimum value should be RT difference = −0.024 s at color difference = 0 (white disk), and that function should rise to RT difference = 0 for the conditions where both target and distractor are on the same side of zero (black curve). The average value of RT difference at color difference = 0 is statistically significantly different from the prediction. (d–f) Analysis of the results of Experiment II. (g–i) Analysis of the results of Experiment III. (d, g) Conventions as in (a). (e, h) Conventions as in (b). (f, i) Conventions as in (c).
opposite direction from our experimental hypothesis. Thus, our result has the statistical power to specifically reject the \(0.024\) s reported by Gilbert et al.

### Experiment II

Many investigators propose the Sapir–Whorf hypothesis as a general theory of visual perception under real-life conditions. Therefore, it is important to ask: Are there experimental conditions under which the result of Gilbert et al. holds? The Munsell color order system used by Gilbert et al. and in Experiment I was based on the appearance of colors (see Nickerson, 1940 for a historical review). Perhaps RT depends on the discriminability of colors instead of their appearance. Experiment II was designed to examine the possible generality of the results of Experiment I by using colors that were approximately equally spaced within CIELAB color space. We chose CIELAB because it was based on the color discrimination data of MacAdam and his colleagues (Wyszecki & Stiles, 1982).

Button-press RT has limited ecological validity: Our forebears did not press buttons. On the other hand, human beings and many other animals do look at stimuli that they can see, and in the case of humans, this looking behavior is often in the form of saccadic eye movements. Therefore, in Experiment II, we presented the stimuli in a saccadic eye movement paradigm and measured the amount of time between the appearance of the stimulus display and the moment the subject’s gaze reached the target stimulus.

The Green–Blue boundary

As in Experiment I, the Green–Blue boundary was measured on each observer using the MOA described above. The stimuli were presented using a CRT video monitor apparatus. The colors were at 35.4 cd/m\(^2\), the target and distractors were separated by approximately 15° in CIELAB color space, and the surrounding field was at 25.2 cd/m\(^2\). Eight monolingual, native English-speaking adults, aged 20–60, served as observers in the main experiment. All were right-handed (5 females) and all had normal color vision (D-15 panel). Except for coauthor AMB, all observers were experimentally naive.

### Visual search

The stimuli for the visual search experiment were presented on a rear projection screen using a video projector (Appendix A). After the observer was fixating a tiny dot in the center of the screen, a 750-ms black annulus (diameter = 2.6° v.a., thickness = 0.1° v.a.) appeared, then was replaced by a ring of 12 colored disks (Figure 1b), which remained in place until the observer responded. The observer’s task was to determine which of the colored disks (the target) had a different color from the others and to fixate it as quickly as possible, while maintaining good accuracy in the choice of disk. The chromaticities of the stimuli are in Appendix A (Figure A1b). The colors were spaced approximately every 4.5° of azimuth in CIELAB color space, but the greenest color (as target or distractor) was paired with fourth greenest color (as distractor or target), and so forth, so the target and distractor colors were always separated by about 14° of azimuth in CIELAB color space (average \(\Delta E = 10.7\)). This strategy ensured that at least one stimulus–distractor color combination would straddle the boundary for each observer, no matter where over the tested range the Green–Blue boundary fell. An eye tracker (Tobii X120) measured the saccadic RT, that is, the elapsed time between the onset of the stimulus and the time when the first saccadic eye movement arrived at the correct target.

### Results of Experiment II

Observers fixated the correct stimulus on 94.5% of trials, and each observer contributed data from between 5 and 11 iterations (median = 9.5) through the stimulus set, depending on observer availability and the proper function of the equipment. A ninth observer’s data were discarded because of technical difficulties with the eye tracker.

Each observer’s RT data set (Figure 5) is presented along with the corresponding Green–Blue boundary value, which is indicated by a black triangle just below his/her
RT data set. According to Kay and Kempton (1984) and Gilbert et al., the Sapir–Whorf hypothesis predicts that this estimate of the fastest RT will occur at the Green–Blue boundary. Contrary to that prediction, visual inspection reveals that the numerical minimum of these RT data sets was not consistently close to the boundary. As in Experiment I, we estimated the color azimuth of the fastest color in each observer’s data as the minimum of a quadratic function fitted to the range of the five RTs indicated by the dashed line on the graph of the average data (Figure 3d). The fastest color (184.97° in CIELAB, SD = 1.67) was reliably greener than the Green–Blue boundary (193.63° in CIELAB, SD = 1.67; paired t7 = 3.40, p = 0.012), and the two values were not reliably positively correlated with one another (r = −0.310, p = 0.228; Figure 3e). The coefficients of those quadratic functions were statistically significantly greater than zero (t7 = 5.96, p = 0.0006), indicating that the data show a local minimum over the range of colors indicated by the dashed line.

Data analysis

As in Experiment I, we examined the linear RVF–LVF RT difference data, normalized along the x-axis so that each subject’s Green–Blue boundary was at color difference = 0 (Figure 3f). The resulting analysis of the 95% confidence intervals around the value at color difference = 0 was y = −0.020058 to y = +0.0214, which excludes the value y = −0.024. Thus, this data set excluded Gilbert’s result.

Experiment III

Perhaps the problem with Experiment II was that the dependent measure was saccadic latency rather than manual button presses, as in Gilbert et al. For example, others have pointed out that an extrageniculostriate pathway and perhaps a geniculostriate pathway via the dorsal stream are probably involved in targeting saccadic eye movements to a certain location in oculocentric visual space. In contrast, the geniculostriate ventral stream probably mediates cognitive decisions such as whether to press the right or left button, according to where the target is located in allocentric visual space. Any Whorfian language effect might be more important for a ventral stream task than for a dorsal stream task, so maybe our failure to replicate Gilbert et al. in Experiment II was due to the task rather than to any fundamental failure of the Whorf hypothesis. Therefore, in Experiment III, we repeated Experiment II using a manual button-press task rather than a saccadic fixation task.

The RT and MOA experiments were performed on the same CRT apparatus as Experiment I (Appendix A). The chromaticities were paired (target and distractor colors) into 5 color pairs, each target and distractor being separated by approximately 15.7° of azimuth in CIELAB color space (average ∆E = 11.5; Figure A1c). The averages of the target and distractor stimulus chromaticities were spaced every 10° around a circle in CIELAB color space. Here, the stimuli were 12 square patches of color (Figure 1a), as in Experiment I. The observer pressed the “S” key with the left hand if the target stimulus appeared to the left of the fixation point and the “L” key with the right hand if the target appeared on the right. We collected 8 iterations through the RT experiment, on 13 color-normal adult observers (D-15 panel; ages 24–62, median = 37; 9 females). The observers also provided MOA measurements of their Green–Blue boundaries.

Results

Observers pressed the correct key on 97.6% of trials. Inspection of the individual RT data (Figure 6) shows no reliable minimum at the Green–Blue boundary. As before, we estimated the color azimuth of the fastest color by fitting a quadratic equation to each observer’s data set over the range of three stimulus combinations indicated by the dashed line near the average data in Figure 3g. The average minimum was at 181.26° in CIELAB, SD = 4.42, and was at a reliably greener color than the MOA boundaries (201.58° in CIELAB, SD = 5.59; paired t13 = 8.701, p < 0.001). Furthermore, there was no relation between the azimuths of the RT minima and the MOA boundary data (r = −0.405, p = 0.067; Figure 3h). As in Experiment II, we examined the coefficients of the quadratic functions that went into that analysis and found that they were reliably positive (t12 = 15.40, p < 0.0001), indicating that the RT functions were reliably U-shaped. We normalized the data as in Experiments I and II, by subtracting the LVF data from the RVF data and subtracting each observer’s Green–Blue boundary from the color azimuth value, for each color azimuth (Figure 3i). The SPSS Mixed-Procedure Analysis revealed that the 95% confidence interval for the difference between the LVF and RVF data at the boundary was y = +0.015 to y = +0.042. Thus, this data set was also inconsistent with the idea that the Whorf hypothesis applies only to the RVF, and the −0.024-s difference between the LVF and RVF RTs at the Green–Blue data was specifically excluded by this data set. Thus, the failure of Experiment II to reveal a consistent effect like what Gilbert reported was not due to our use of saccadic RT as the dependent measure.

This experiment was designed using button-press RT to determine whether the results of Experiment II were due to the use of saccadic RT as a dependent measure. The results of Experiments II and III resembled each other.
closely. Three observers served in both experiments (coauthor AMB, KTN, and BUI, indicated by the symbols in Figures 5 and 6), so it is of interest to compare the results of those observers across the two methodologies (Figure 7a–7c). Saccadic RT was generally faster than button-press RT. This suggested that either the time required to execute the motor response was longer for the button-press task than the eye movement task or else the choice between the 12 colors in Experiment II took less time than the choice between the two response keys in Experiment III. In any case, both of these tasks are choice RTs, and the RTs are always longer than the classic saccadic or button-press latency values for simple RTs. Furthermore, each observer’s RT function had a consistent shape across the two experiments, which suggested that the shape was governed by sensory and perceptual factors that were common to the two experiments rather than random variation or constant factors related to the decision or to the generation of the response itself.

**Experiment IV**

We also ran an intensive series of saccadic RT measurements on three observers to explore the impact of two decisions we made in Experiment II: the decision to use a dark rather than a light background and the decision to have our observers look at the target, which did not require an explicit left–right decision before responding. These three observers were tested in a $2 \times 2$ experimental design. The first factor was a light vs. a dark surrounding field. The second factor was task: The observers either looked at the target, as in Experiment II, or they looked at a “button” dot affixed to the right or left side of the screen (Figure 1b) according to whether the target appeared on the right or left half of the display. The apparatus for the RT experiment was the same rear projection video screen as we used for the RT data in Experiment II. We used the MOA to measure the Green–Blue boundary of each observer, as in Experiment II, only using the rear projection screen apparatus. The pairs of target and distractor stimuli were separated by about $15.9^\circ$ in CIELAB color space (average $\Delta E = 17.2$, Appendix A).

**Results**

A trial was accepted as correct if the first saccadic eye movement arrived within an area of interest around the correct target in the look-at-the-target condition and around the correct “button” on the correct side in the look-at-the-button condition. By these criteria, performance was 88% correct in the look-at-the-target condition and 90.7% correct in the look-at-the-button condition. The color at which the fastest RT occurred was not reliably...
related to the Green–Blue boundary, as can be seen in the individual data sets (Figure 8). As in Experiments I–III, we fitted descriptive quadratic equations to the RT data sets over the range of five data points indicated by the dashed lines at the top of Figure 8 and extracted an estimate of the local RT minimum for each subject under each condition within that range. The average of all the RT minima in the experiment was at 178.38° in CIELAB, SD = 5.33, which was greener than the MOA settings (199.97° in CIELAB, SD = 5.07, paired t11 = 2.91, p = 0.014). As before, there was no clear relation between the RT minima extracted in this way and the Green–Blue boundaries measured using MOA (Figure 9f; overall r = 0.015, p = 0.4813). The average value of the coefficients of the quadratic term was reliably greater than zero (t11 = 9.494, p < 0.0001), indicating that the data, when pooled across all the conditions, were U-shaped.

Because of the small number of observers in Experiment IV, further statistical analysis is unwarranted. By inspection, we note that one of our three observers (coauthor KMG, indicated by the asterisk) showed extra-fast RT at the Green–Blue boundary in one of the four conditions. However, in that case, the RT in the RVF was not faster than in the LVF. We processed the results of each observer using the methods described for Experiments I–III. The RT difference values are shown as a function of the color difference values in Figures 9a–9d. Several of the data sets fell below the Y = 0 line, indicating that RT was faster in the RVF than in the LVF. However, there was no clear trend for the data in any of the conditions to be a U-shaped function with its minimum at −0.024 s, as would be predicted by the results of Gilbert et al. Thus, none of the conditions produced an effect similar to that of Gilbert et al. The lack of such an effect suggests that our choices of task and surrounding field lightness were not crucial to our failure to find a Whorfian color boundary effect in the RVF only. Subject and coauthor KMG served in both Experiment IV (saccadic RT) and Experiment III (button-press RT; Figure 7d). As for the other subjects in Figure 7, his saccadic RT was slightly faster than the button-press RT, but the overall shapes of the two functions were similar.

Figure 7. (a–d) RT data from four individual observers who served in both button-press RT and saccadic RT experiments. Subjects AMB, KTN, and BUI served in Experiments II and III; subject KMG served in Experiments III and IV. RT for the saccadic eye movements to one of 12 color samples (black disks) was reliably faster than for the button press of one of two response keys (white disks), and the shape of the RT function for each subject was similar across the two tasks. (e, f) Average saccadic RT data from Experiment IV. RT for the look-at-the-target task (black symbols) was reliably faster than for the look-at-the-button task (white symbols).

Figure 8. Three observers’ data from Experiment IV. White symbols: RVF; black symbols: LVF; circles: look at target, dark surround; a, look at target, light surround; b, left–right, light surround; c, look at target, dark surround; d left–right, dark surround. Upright black triangles: Green–Blue boundaries. Displacement parameters for the three subjects were 0.0 (◦), 0.25 s (†), and 0.35 s (‡).
These data allowed us to examine, qualitatively, the effects of the look-at-the-target vs. the look-at-the button instructions. The data (Figures 7e and 7f), averaged across three subjects and both sides of the visual field, revealed that the look-at-the target data were consistently faster (average difference = 0.038 s). The results are similar to those obtained when look-at-the-target saccadic RT data are compared to the right–left button-press RT data. Apparently, the “choice” in these choice RT tasks had an important effect on RT.

Experiment V

While Experiments I–IV revealed no reliably extra-fast RT at the Green–Blue boundary specific to the RVF, as Gilbert et al. reported, the data sets did show a large range of RT values. What was this variability due to? One possibility is that the color combinations might not all look equally different to the observer, even though they were approximately equally separated in the Munsell color order system or the CIELAB color space. To investigate this possibility, we measured the perceived differences between the stimuli from Experiments II and IV using Maximum Likelihood Difference Scaling (MLDS), a modern psychophysical scaling technique (Knoblauch & Maloney, 2008; Maloney & Yang, 2003).

Methods for Experiment V

The observers were the same people who served in Experiments II and IV, and the apparatus was the same as was used for the MOA data (the CRT for the observers from Experiment II and the rear projection apparatus for the observers from Experiment IV). For the observers from Experiment II, the stimulus array was two pair disks (Figure 1c). Each top or bottom pair of colors was separated by one, two, or three chromaticity steps, from a set of 11 equally spaced colors (stimulus chromaticities are in Appendix A). The steps were chosen so that two steps in the MLDS experiment corresponded approximately to the difference between the target and distractors in the RT experiment. The surrounding field was dark gray. For the observers from Experiment IV, the surrounding field was either dark or light, and the targets were grouped on the right or left half of the visual field (Figure 1d). There were 10 equally spaced colors, and two steps in the MLDS experiment corresponded approximately to the difference between the target and distractors in the RT experiment.

On each MLDS trial, the observer judged which of the two pairs of colors was the more dissimilar: the top pair or the bottom pair. This was a non-verbal judgment in the sense that no color terms were needed. In the case of the observers from Experiment II, the data were collected before the MOA and RT data, so the experimentally naive observers did not know that they would be
asked to name or judge any colors. Only coauthor AMB knew what hypotheses were tested. Each observer from Experiment II contributed one run through the MLDS stimulus set. In the case of the observers from Experiment IV, the MLDS measurements were made after the RT data and MOA data were collected; only coauthor KMG knew that this was a study of categorical perception at the Green–Blue boundary. Each of the three observers from Experiment IV contributed 10 runs through the stimulus set (five runs in the LFV and five in the RVF).

Results

The MLDS algorithms of Knoblauch and Maloney (2008) generated a curve of “Psy” (∆ψ), the scaled magnitude of the stimuli, normalized to the range (0, ..., 1), as a function of the stimulus values, which were the color angles of the stimuli in CIELAB color space. Figure 10a shows the average ∆ψ curves for the observers from Experiment II. The difference between the ∆ψ values of two color angles is the scaled perceptual difference between them, delta Psy (∆ψ; e.g., the red brace in Figure 10a). In our case, the two color angles were separated by two steps along the ∆ψ function, because those were separated by approximately the same amount (in azimuth of CIELAB) as the targets and distractors in Experiment II. This analysis generated a graph (Figure 10b) that related stimulus chromaticity to stimulus appearance under our particular conditions: The x-axis was the azimuth in CIELAB color space (averaged across the two color angles), and the y-axis was the scaled perceived differences between target and distractor stimuli (∆ψ values). For example, the two colors indicated by the red brace in Figure 10a were subtracted to obtain the value of ∆ψ at the tip of the arrow in Figure 10b. If categorical perception had occurred, we would expect a locally high

Figure 10. MLDS data and their fits to the RT data. (a–c) Data from the observers from Experiment II, using stimulus configuration from Figure 1b. Black triangles, MOA, Green–Blue boundaries. (a) Squares, MLDS ∆ψ data. (b) Diamonds, ∆ψ data derived from (a). (c) The reciprocal of the ∆ψ data (line) was fitted to the RT data of Experiment II using Equation 1 (circles). (d–i) Data from the observers from Experiment IV; red and black solid lines, RVF and LVF, respectively. (d–f) Dark surrounding field. (g–i) Light surrounding field. (d, g) Squares, MLDS ∆ψ data; solid lines, point-to-point data. (e, h) Diamonds, ∆ψ data; solid lines, point-to-point data. (f, i) The reciprocals of the ∆ψ data (solid lines) were fitted to the RT data of Experiment IV using Equation 1 (white circles, RVF; black circles, LVF). White triangles and dashed lines throughout the predicted curves for the RVF, taken from the LVF data, but assuming a 0.024-s category effect at the Green–Blue boundary (f, i).
value of $\Delta \Psi$ at the category boundary, where a green stimulus was compared to a blue stimulus, and lower $\Delta \Psi$ values above and below the boundary (Figure 11a; see Harnad, 1987, for a review). In contrast, if the colors are not perceived categorically, there will be no locally high value at the putative category boundary (Figures 11b–11d).

**Discussion**

On the hypothesis that target stimuli are easier and faster to detect when they are more different from their distractors, and slower and harder to detect when they are more similar, one might suppose that the MLDS and the RT data might be related to one another. To explore this possibility, we predicted the RT data from the reciprocal of the MLDS data:

$$RT(x) = RT_{\text{min}} + k \cdot \left( \frac{1}{\Delta \Psi(x)} - \frac{1}{\Delta \Psi(x)_{\text{min}}} \right), \quad (1)$$

where $\Delta \Psi(x)$ is the MLDS-scaled difference between the two colors of the color combination (e.g., Figure 10a), $x$ is the average of those two color angles, $RT(x)$ is the RT at $x$, and $RT_{\text{min}}$ is the minimum RT for the data set. The average predictions of Equation 1 for the RT results of Experiment II appear as the black line in Figure 10c. For example, the two values of $\Psi$ indicated in red in Figure 10a produced the value of $\Delta \Psi$ indicated by the arrow in Figure 10b and predicted the value of RT at the tip of the arrow in Figure 10c.

The predicted RTs from the much more extensive individual MLDS $\Psi$ data on the three observers from Experiment IV are in Figure 12. The fits are good, suggesting that RT can be understood from the perceived differences between the targets and the distractors, regardless of which color categories they come from.

The average $\Psi$ data for the RVF and LVF on the observers from Experiment IV were similar to each other and so were the average $\Delta \Psi$ functions for the RVF and LVF (compare the black and white squares in Figures 10d and 10g and the black and white diamonds in Figures 10e and 10h). Thus, the MLDS data did not suggest any categorical effect restricted to the RVF. However, how big would the effect be, if the RVF RT data showed a dip of 0.024 s near the Green–Blue boundary? To answer that question, we adjusted the MLDS function from the LVF (white triangles and dotted lines in Figures 10f and 10i) to predict a dip of 0.024 s in the predicted RT function (white triangles in Figures 10f and 10i) near the average Green–Blue boundary. The intermediate step in that prediction was the predicted RVF $\Delta \Psi$ function (white triangles in Figures 10e and 10h). That function shows a prominent maximum of the kind illustrated in Figure 11a. Whereas the predicted changes to the RT data (Figures 10f and 10i) and the MLDS $\Psi$ data (Figures 10d and 10g) are subtle, the predicted effects on the $\Delta \Psi$ functions are large and are clearly ruled out by the RVF data (white diamonds, with error bars smaller than the data points, are well clear of the large deviation shown by the white triangles). Thus, the results of Experiment V provide evidence against a categorical effect in the perceived differences between colors that straddle the Green–Blue boundary.

The lack of an RVF–LVF difference in Figures 10d–10i is perhaps not surprising, because the left and right cerebral hemispheres are extensively connected. A Wohrfian effect might cause categorically different stimuli presented to the RVF to be identified sooner because of the primary visual projection from the RVF to the left
categorical effect might produce RT that is faster than the null hypothesis prediction but not necessarily faster than all the colors in the data set (Figure 13a). The Green–Blue boundary would not be the fastest RT (the fastest color might be elsewhere, white triangle in Figure 13a), but it would be the color where the RT data fall below the null hypothesis prediction. The categorical boundary effect would be evident in the error of prediction, which would show a localized minimum (Figure 13b). The minimum RT values for the observers in Experiments I–IV were consistent with this possibility, because those values were all close to 185° in CIELAB but were not closely related to the Green–Blue boundaries (Figures 2, 3b, 3e, 3h, 5, 6, and 8). Therefore, it was especially important to establish a null hypothesis to evaluate our data, to determine whether the scenario illustrated in Figure 13 applies here.

The choice of color space is a null hypothesis

Investigators in the tradition of Kay and Kempton (1984) generally attempt to assure that their results are due to a perceptual rather than a sensory difference between their stimuli by choosing stimuli that are separated by a constant distance in some presumably uniform metric color space. However, the rigorous assumption of uniformity of any color space is unwarranted. Uniform chromaticity spaces differ from one another in several ways, including in their design criteria. For example, the color samples that define the Munsell color order system were chosen to be perceptually uniformly spaced but not necessarily to be uniform with respect to sensory color discrimination. Indeed, stimuli that are separated by constant distance in Munsell space are not generally equally discriminable from one another, as specified by CIELAB (e.g., Kuehni, 1999). More generally, the separation of two colors in JND units (by Fechnerian scaling) is known to be an unreliable guide to their perceived difference (Wyszecki, 1972), and the perceptual uniformity of the Munsell color system, even with respect to perceived hue, saturation, and value, is at best an approximation (Indow, 1988). The spacing of the colors in CIELAB and CIELUV uniform chromaticity spaces was intended to represent approximately constant-sized JNDs in all locations and in all directions of color space, based on MacAdam’s ellipses (Wyszecki & Stiles, 1982), but even these spaces are both non-uniform and non-isotropic (e.g., Figure 5.6 in Shevell, 2003 and Figure 5.4.1 in Wyszecki & Stiles, 1982). The designs of CIELUV and CIELAB were heavily influenced by the desire for the uniform chromaticity spaces to be computationally simple transformations of CIE color matching functions. This means that the discriminability of equally spaced colors can vary substantially, even when they are close together. The problem is even more severe when the “within-category” stimuli are located in a different region.
of color space than the “between-category” stimuli (e.g., Daoutis et al., 2006; data reanalyzed and discussed in depth in Drivonikou, Davies, Franklin, & Taylor, 2007). Furthermore, all color spaces, including the Munsell color system, CIELAB, and CIELUV, were designed based on average data, with considerable smoothing, so they will not apply to any single observer. This makes them inadequate benchmarks against which to assess categorical perception of individuals. Finally, the general model that RT should be monotonic with separation in color space is incorrect in the limit, because there is a maximum separation in color space beyond which RT is fast and constant (Nagy & Sanchez, 1990). In view of the modest sizes of some of the effects reported in this literature, it is striking that so little attention has been paid to the metric that quantifies the differences between the stimuli.

It is now clear that simply equating the separations of the stimuli along some psychological continuum is sometimes not sufficient to produce constant RT in a visual search experiment. In previous work (Lindsey et al., 2010), we and our collaborators studied visual search using colors that were carefully adjusted for their psychological properties: Their colorimetric purity was adjusted for constant saturation, and the subjective separations in color between targets and distractors were carefully controlled. In spite of these controls, we found that RT varied drastically with hue. In contrast, RT was predicted very well by a standard “low-level” color-opponent model, which was based on the responses of the well-understood LM and S channels. Thus, a failure of stimuli that have certain carefully established subjective qualities to behave as expected does not necessarily mean that some higher level process must be at work: On the contrary, a low-level sensory model might very well account for the results.

The standard color model as the null hypothesis

To get around these difficulties and to test directly the hypothesis that the perception of colors is categorical, we need a model of RT for colors. Ideally, this model will be a purely sensory model and it will make specific predictions that can be falsified by the results of perceptual studies such as the study of Gilbert et al. What should the RT be, in the conditions of a particular experiment, under Whorfian and non-Whorfian assumptions?

Our null hypothesis was a standard color-opponent model of discrimination between colors. That is, we assumed that when the target and distractor colors are hard to discriminate, RT should be slow, and if the colors are easy to discriminate, RT should be fast. Therefore, we predicted the results of Experiments I–IV, using a standard color-opponent model (see Lindsey et al., 2010, Supplement and Appendix C below, for further details). Briefly, we calculated the coordinates of our stimuli in MacLeod–Boynton color space, which has axes directly related to L–M excitation and S-cone excitation, at constant luminance (MacLeod & Boynton, 1979). Then, we calculated the difference in excitation presented to the L–M and S-cone channels by the colorimetric difference between target and distractors. Our predicted RT was a function of these differences:

$$
RT(x) = a + \left( b \cdot |\Delta LM(x)|^{-1} \right)^4 + \left( c \cdot |\Delta S(x)|^{-1} \cdot g(x) \right)^4 \right)^{1/4},
$$

where

$$
g(x) = \frac{50 + 0.14 \cdot S(x)}{50 + 0.14 \cdot S_{min}}.
$$

In Equation 2, $x$ is the color azimuth in CIELAB, $a$ is the minimum possible RT, $b$ and $c$ are scalars on the overall contribution of the L–M and S components (values of $a$, $b$, and $c$ are in Appendix C), and $|\Delta LM(x)|$ and $|\Delta S(x)|$ are the absolute values of the differences in L–M and S excitation of the target and distractors. $g(x)$ is a factor from Boynton and Kambe (1980) that takes the increment threshold function for the S cones into account and is based on an independent experiment (Lindsey et al., 2010, Supplement). The exponent of 4 is the Quick pooling formula, which we use instead of the ArgMin function of Lindsey et al.

The fits of Equation 2 appear in the left-hand panels of Figures 14 and 15. The green lines in Figures 14a–14c and 15a–15d are the contribution of the L–M channel, the blue lines are the contribution of the S channel, and the red lines are the fits of Equation 2, with parameters $b$ and $c$ fitted by a least-squares criterion. The minimum RT (parameter $a$) was held constant at 0.4 s for the button-press RT experiments (Experiments I and III) and 0.26 for the saccadic RT experiments (Experiments II and IV), consistent with the observation that the saccadic RT functions are consistently faster overall but not very different in shape across the two methodologies (Figures 7a–7d). The L–M channel increases without bound in Figures 14a and 15a–15d in the neighborhood of the tritan point where $L = M$ and their difference approaches zero. The fits to the RT data from Experiments I–III (Figures 14d–14f) and IV (Figures 15e–15h) are reasonably good. The errors of prediction are the red curves in the right-hand panels of Figures 14 and 15. If there were categorical perception at the Green–Blue boundary (black triangles), there would be a local minimum in the error function at that color. There is little evidence of such a local minimum.

This simple, low-level model works reasonably well, which suggests that the speed with which subjects can do...
the visual search task is largely controlled by the strength
of the sensory color signals available to mediate the
response. Particularly, there is no reliable localized
negative peak in the error graphs corresponding to the
Green–Blue boundary. The success of these fits indicates
that the standard color-opponent model of color discrim-
ation is sufficient to account for these RT results,
including the minimum values in the neighborhood of
185° of azimuth within CIELAB, without invoking the
complication of categorical perception. We suspect that
the minima reported by other authors may also be the
result of low-level factors that control the sensitivity of
the eye to changes of stimulus chromaticity.

General discussion

In their 2006 paper, Gilbert et al. proposed that the
Sapir–Whorf hypothesis applies to the right visual field
(RVF) but not the left visual field (LVF). This is an
attractive idea, because it seems likely that a Whorfian
influence of language on perception would be stronger
when the visual stimulus is presented to the cerebral
hemisphere where the language centers reside. Gilbert et
al. reported a lateralized Whorfian effect in visual search:
Right-handed observers were about 0.024 s faster to find a
target when it was from a different color category from its
distractors (e.g., a blue target among green distractors)
than when it was from the same category (e.g., a blue
target among blue distractors) but only when the target
was presented in the RVF. When the target fell within the
LVF, the within-category and across-category stimulus
combinations produced similar response times (RTs). The
RT minimum was held to be evidence of categorical
perception because it coincided with the Green–Blue
boundary measured by the Method of Constant Stimuli.
However, that experiment provided no quantitative anal-
ysis substantiating the null hypothesis that no RT mini-
mum should occur in the absence of categorical
perception, and no alternative explanation of the location
of the RT minimum in color space was discussed or
ruled out.

Figure 14. The RT data from (a–c) Experiments I, II, and III were
fitted using Equations 2 and 3. Green lines, L–M contribution; blue
lines, S contribution; red lines, the full model. (d–f) Errors of
prediction from Equation 2 (red lines) and from the MLDS fits in
the case of Experiment II (black line in (e)). Black triangles: the
average Green–Blue boundaries; green arrow: the L–M contribu-
tion goes up to infinity at the tritan colors where L = M. The errors
of prediction do not show a pronounced local minimum at the
Green–Blue boundary.

Figure 15. (Left) The average RT data from Experiment IV, fitted
using Equations 2 and 3. (Right) Errors of prediction from
Equation 2 (red lines) and the MLDS fits. Conventions as in
Figure 14.
This basic result has been the subject of 17 empirical studies that appeared since Gilbert et al.’s paper was published; of these, seven articles reported psychophysical data (Drivonikou, Kay et al., 2007; Franklin, Drivonikou, Bevis et al., 2008; Franklin, Drivonikou, Clifford et al., 2008; Liu et al., 2009; Roberson, Pak, & Hanley, 2008; Siok et al., 2009; Zhou et al., 2010). All of these papers reported a minimum RT value in the neighborhood of 185° of color azimuth, which was near the Green–Blue boundary measured using MCS (see Appendix B, below). Four of the articles reported a statistically significant RVF–LVF RT difference in the size of that effect (Drivonikou, Kay et al., 2007; Franklin, Drivonikou, Bevis et al., 2008; Franklin, Drivonikou, Clifford et al., 2008; Roberson et al., 2008), and three reported no statistically significant difference (Liu et al., 2009; Siok et al., 2009; Zhou et al., 2010). The RT experiments we report here were designed to replicate or refute that important result and to find the experimental conditions under which it holds.

We were not able to replicate the RT result of Gilbert et al., using either button-press RT or saccadic RT as a dependent measure, using either their own stimuli or stimuli chosen within the CIELAB color space, or using stimuli with light or dark surrounding fields. Like many other workers in this field, we did find reliable RT minima in the neighborhood of 185° of color azimuth within CIELAB in all our data sets. However, these minima were not generally obtained with the between-category stimuli, in either visual field, and a −0.024-s RVF–LVF difference in cross-category RT, the magnitude reported by Gilbert et al., was not consistent with our data.

In contrast to the lack of evidence for a reliable Whorfian effect in any of our data sets, our two modeling efforts worked well. Most directly, the scaled reciprocal of the ΔV data from the MLDS experiments fit the RT data well. This fit is intuitively appropriate, but it has the difficulty that if there were a perceptual discontinuity at the Green–Blue boundary, that discontinuity could affect both the RT and the MLDS data similarly. Therefore, both data sets might have a local minimum at the Green–Blue boundary, but no discontinuity in the fit between the two data sets is to be expected. In a similar vein, the use of the Munsell space for an experiment like that of Gilbert is problematical. If there were indeed a Whorfian effect of language on perception, that perceptual effect should have already adjusted the Munsell color space to be uniform with respect to perceived color differences, so no category effect should be found.

It is more convincing that the standard model of “early” color vision fit the data reasonably well (red lines in Figures 14 and 15). The important feature of these models is that they provide reasonable fits to the full range of RT data, from the greenest to the bluest stimuli, with no assumptions about the uniformity of the color space within which they were chosen and without recourse to a Whorfian element or any other high-level cognitive explanation. Indeed, the three estimates of the differences of the colors, i.e., the RT data, the perceived differences from the MLDS experiments, and the predicted sensory differences from color theory, are all very similar. The errors of prediction obtained when the RT data are predicted from the perceived differences and the sensory differences are very similar (compare the black and red lines in the right-hand panels of Figures 14 and 15). This leads us to conclude that the Whorfian hypothesis is not a necessary component of any complete theory of visual perception of the colors in these experiments.

**Conclusion**

The results of the experiments reported here call into question the traditional use of color perception at the Green–Blue boundary as a paradigm for studying the possible correspondence between perceptual categories and color names. Although colors are necessarily categorized when they are named (otherwise, one would need a distinct name for each discriminable color), they are apparently not categorized when they are perceived, at least not under the particular experimental conditions we examined. Of course, it is not appropriate to generalize beyond these data and analyses to speculate whether visual perception is ever categorical. However, these results do challenge the status of the Sapir–Whorf hypothesis as a general theory of visual perception. Furthermore, the use of perceived hue and color terms as a way of studying the possible relation between perceptual and linguistic categories should be reexamined critically, at least for reaction time experiments involving stimuli in the neighborhood of the Green–Blue boundary.

**Appendix A**

**Stimulus parameters for Experiment I**

**Apparatus**

Stimuli were presented using a Mitsubishi Diamond-Pro 9TTFM cathode-ray tube (CRT) video computer monitor. RT responses were entered by means of the computer keyboard. The observer moved the computer mouse to manipulate the colors in the MOA phase of the experiment.

**Spatial parameters**

For the RT experiment, the target and distractor stimuli were twelve 1.5° × 1.5° v.a. squares, presented with their
centers equally spaced around a 5° v.a. radius circle. For the Method of Adjustment (MOA), the stimulus in Position 1 (Figure 1a) was the “target” color and the other 11 stimuli were the “distractors.”

**Chromaticities**

The chromaticities of the stimuli appear above (Figure A1a), in CIELAB units, calculated from the calibrated xyY coordinates (Pritchard PR-670 SpectraScan spectrophotometer) with a white point of (0.310, 0.316, 80). The luminance of the colors was 39.5 cd/m², and the luminance of the surrounding gray field was 30.3 cd/m². For the MOA, the chromaticities of the stimuli moved along the curve defined by the white disks, with the target and distractor separated by 15° of azimuth in CIELAB.

**Stimulus parameters for Experiment II**

**Apparatus**

For the RT experiment, stimuli were presented on a high-diffusion rear projection screen by a DILA video projector (JVC DLA-M2000L). RT responses were in the form of eye movements, which were recorded by a Tobii X120 Eye Tracker (Stockholm, Sweden). In the MOA phase of the experiment, a ViewSonic P815 CRT was used to present the stimuli, and the observer moved the computer mouse to manipulate the colors.

**Spatial parameters**

For the RT experiment, the target and distractor stimuli were 3.5° v.a. diameter disks, presented with their centers equally spaced around a 15° v.a. radius circle. For the MOA measurements, the target was in Position 2 (Figure 1b), and the other 11 stimuli were distractors.

**Chromaticities**

The chromaticities of the visual search stimuli appear as the white disks in Figure A1b, obtained from the calibrated (Pritchard PR-670 SpectraScan) xyY values using a white point at (0.310, 0.316, 135). The colors were 65.23 cd/m², and the surrounding field was 49.1 cd/m². The colors were separated by an approximately constant 14° of azimuth in CIELAB color space and an average $E$ of 10.7 (range: 9.77–11.84). For the MOA, the stimuli were adjusted along the contour described by the white disks, with a constant separation of 15° of azimuth in CIELAB.

**Stimulus parameters for Experiment III**

**Apparatus**

The RT and MOA apparatus was the same as for Experiment I.

**Spatial parameters**

Target and distractor stimuli were 3.2° × 3.2° v.a. squares, presented with their centers equally spaced around a 10° v.a. radius circle. For the Method of Adjustment (MOA), the stimulus in Position 1 was the “target” color and the other 11 stimuli were the “distractors.”

**Chromaticities**

The calibrated chromaticities of the stimuli appear in Figure A1c, which were calculated from the calibrated...
The x/y/Y coordinates using a white point of (0.310, 0.316, 80). The colors were 39.55 cd/m², and the surrounding field was 30.25 cd/m². The colors were separated by about 15.7° of azimuth in CIELAB or 11.5 ΔE units (range: 11.1–11.97). For the MOA, the stimuli were adjusted along the contour described by the white disks, with a constant separation of 15° of azimuth in CIELAB.

Stimulus parameters for Experiment IV

Apparatus

Same as for Experiment II (for both RT and MOA).

Spatial parameters

The search stimuli and the stimuli for the MOA measurements were the same as for Experiment II. Only in the “look-at-the-button” conditions, small paper “buttons” were affixed to the rear projection screen midway between the fixation target and the left-hand and right-hand color disks (Figure 1b); the “buttons” were absent during the “look-at-the-stimulus” experiments. The observer looked at the appropriate “button” on each trial to indicate his/her choice of the right-side vs. left-side location of the target.

Chromaticities

The chromaticities for the RT experiments appear in Figure A1d, obtained from the calibrated x/y/Y values using white point at (0.310, 0.316, 135). The colors were 55 cd/m² and were separated by about 25.3° of azimuth in CIELAB; average ΔE was 19.1 (range: 17.8–20.8). The surrounding field was 41.3 cd/m² in the “dark surround” condition and 96.3 cd/m² in the “light surround” condition.

Stimulus parameters for Experiment V

Apparatus

The same as for the corresponding MOA experiment (a CRT for the observers from Experiment II and a rear projection apparatus for the observers from Experiment IV). The observers responded by pressing keys on the computer mouse.

Spatial parameters

As in Experiments II and IV, the disk diameters were 3.5° v.a. For the observers from Experiment II, the two pairs of colored disks that were judged were in positions 12 and 1, and 6 and 7, respectively (Figure 1c). For the observers from Experiment IV, the “near” disks on the right-hand side of the display were centered 3.5° v.a. to the right of the midline, and the “far” disks were centered 7° v.a. to the right of the midline (Figure 1d). The center-to-center vertical separation was 7° v.a. The stimuli on the left-hand side of the display were mirror images of those on the right.

Appendix B

Measuring the Green–Blue boundary

In Experiments I–IV, we measured each observer’s color category boundaries directly using the Method of Adjustment (MOA). While the Method of Constant Stimuli (MCS) enjoys a better reputation among psychophysicists, it is unsuitable for this purpose because, as we verified in a pilot study, the results of an MCS experiment are greatly influenced by the domain of stimuli being judged.

The Method of Constant Stimuli (MCS)

In the MCS pilot study, observers named single colors, using stimuli with chromaticities and luminances similar to those used in Experiment II. The colors were presented in a randomized block design. Interpolation yielded the Green–Blue boundary, i.e., the color angle that was named “green” and “blue” equally often. We varied the midpoint of the domain of stimuli being tested from block to block. The function relating the midpoint of the test domain to the Green–Blue boundary had a slope of 0.318 to 0.332, and the MCS estimates of the Green–Blue boundary covered a range of 20° to 24° of color azimuth (Figure B1, black triangles). In contrast, the Green–Blue boundary was constant when the task was the MOA, performed using the same method as the main experiment (Figure B1, white triangles).

The Method of Adjustment (MOA)

We also collected MOA data for comparison to the MLDS data. In the MOA experiment, the two colors...
Maintained a hue angle separation of 15° of azimuth in CIELAB color space, as the observer adjusted their average chromaticity continuously along a constant-eccentricity contour in CIELAB. Each MOA trial began with a new, randomly chosen start point within the range of possible stimulus hues. While fixating a dot in the center of the screen, the observer adjusted the chromaticity of the stimuli until the “target” disk (disk 1 in Figure 1b) was just green and the other eleven disks were just blue, or vice versa (see Webster & Mollon, 1993 for a similar approach to flicker). If this held for a range of possible settings, the observer adjusted the setting until the bluer stimulus looked as blue as the greener stimulus looked green and the Green–Blue boundary bisected the interval. The observer made 10 settings with square 1 greener than the others and 10 settings with square 1 bluer than the others. The observer’s Green–Blue boundary was taken as the average of 8 settings from each set, trimming the bluest and the greenest settings in each case.

Figure B1. The effects of testing range on the estimated Green–Blue boundary. Black triangles, Method of Constant Stimuli. White triangles: the Method of Adjustment, using the two-stimulus method described in the text. The Method of Constant Stimuli produced clear variation in the measured boundary (**p < 0.0005, in each case), whereas the MOA was more reliable (p > 0.25 in each case). However, the tendency to follow the range was not perfect, as the slope = 1 hypothesis (dashed lines) is also rejected in each case.

(square I vs. squares 2–11 in Figure 1b) maintained a hue angle separation of 15° of azimuth in CIELAB color space, as the observer adjusted their average chromaticity continuously along a constant-eccentricity contour in CIELAB. Each MOA trial began with a new, randomly chosen start point within the range of possible stimulus hues. While fixating a dot in the center of the screen, the observer adjusted the chromaticity of the stimuli until the “target” disk (disk 1 in Figure 1b) was just green and the other eleven disks were just blue, or vice versa (see Webster & Mollon, 1993 for a similar approach to flicker). If this held for a range of possible settings, the observer adjusted the setting until the bluer stimulus looked as blue as the greener stimulus looked green and the Green–Blue boundary bisected the interval. The observer made 10 settings with square 1 greener than the others and 10 settings with square 1 bluer than the others. The observer’s Green–Blue boundary was taken as the average of 8 settings from each set, trimming the bluest and the greenest settings in each case.

Figure B1 shows the results of this experiment. The measured Green–Blue boundary was highly correlated with the midpoint of the testing range in the case of the MCS measurements and covered a range of as much as 24° of azimuth in CIELAB. In contrast, there was no effect of the range of colors available on the Green–Blue boundary when the MOA was used. Therefore, all the measurements of the Green–Blue boundary in the main experiments were collected using the MOA.

### Appendix C

The parameters of the model fits in the Color-theoretical discussion section

Tables C1 and C2.

### Acknowledgments

This work was supported by the National Eye Institute, National Institutes of Health (R21-EY018321 and R21-EY018321-0251) and by the Ohio Lions Eye Research Institute. We are grateful to Ms. Heather Shamp, Ms. Renee Rambeau, and all our observers for their assistance in collecting the data and to Dr. Loraine Sinnott for statistical consultation.
Commercial relationships: none.
Corresponding author: Angela M. Brown.
Email: brown.112@osu.edu.
Address: College of Optometry, The Ohio State University,
Fry Hall, 338 W 10th Avenue, Columbus, OH 43210-1240,
USA.

Footnote

1Coauthor KMG was aware that this was a study of
color categorical perception, but he was not aware of the
hypotheses being tested at the time the RT measurements
were being made. His many intellectual contributions to
this project occurred after he had served as an observer in
Experiments II and IV.

References

Boynton, R. M., & Kambe, N. (1980). Chromatic differ-
ence steps at moderate size measured along theo-
retically critical axes. Color Research and Application,
5, 13–23.

Daoutis, C. A., Franklin, A., Riddett, A., Clifford, A., &
British Journal of Developmental Psychology, 24,
373–400.

Drivonikou, G. V., Davies, I., Franklin, A., & Taylor, C.
(2007). Lateralisation of colour categorical percep-

Drivonikou, G. V., Kay, P., Regier, T., Ivry, R. B.,
evidence that Whorfian effects are stronger in the
right visual field than the left. Proceedings of the Na-
cional Academy of Sciences of the United States of
America, 104, 1097–1102.

Franklin, A., Drivonikou, G. V., Bevis, L., Davies, I. R. L.,
color is lateralized to the right hemisphere in infants,
but to the left hemisphere in adults. Proceedings of the National Academy of Sciences of the United States of
America, 105, 3221–3225.

Franklin, A., Drivonikou, G. V., Clifford, A., Kay, P.,
categorical perception of color changes with color
term acquisition. Proceedings of the National Academy of Sciences of the United States of America, 105,
18221–18225.

Whorf hypothesis is supported in the right visual field
but not the left. Proceedings of the National Academy
of Sciences of the United States of America, 103,
489–494.

cognitive aspects of categorical perception: A critical
overview. In S. Harnad (Ed.), Categorical perception:
The groundwork of cognition (pp. 1–28). Cambridge,
UK: Cambridge University Press.

Indow, T. (1988). Multidimensional studies of Munsell-

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R.
for the Study of Language and Information.


Knoblauch, K., & Maloney, L. T. (2008). MLDS:
Maximum likelihood difference scaling in R. Journal

Kuehni, R. G. (1999). Hue scale adjustment derived from
the Munsell system. Color Research and Application,
24, 33–37.

Lindsey, D. T., & Brown, A. M. (2002). Color naming and
the phototoxic effects of sunlight on the eye. Psycholog-

Lindsey, D. T., Brown, A. M., Reijnjen, E., Rich, A. N.,
channels, not color appearance or color categories,
guide visual search for desaturated color targets.
Psychological Science, 21, 1208–1214.

Liu, Q., Li, H., Campos, J. L., Wang, Q., Zhang, Y.,
Qiu, J., et al. (2009). The N2pc component in ERP and

MacLeod, D. I. A., & Boynton, R. M. (1979). Chroma-
ticity diagram showing cone excitation by stimuli of
equal luminance. Journal of the Optical Society of
America, 69, 1183–1186.

likelihood difference scaling. Journal of Vision,

differences determined with a visual search task.

Nickerson, D. (1940). History of the Munsell color system
and its scientific application. Journal of the Optical
Society of America, 30, 575–586.


