Categorical facilitation with equally discriminable colors

Christoph Witzel
Laboratoire Psychologie de la Perception, Université Paris Descartes, Paris, France

Karl R. Gegenfurtner
Department of Psychology, Justus Liebig University, Giessen, Germany

This study investigates the impact of language on color perception. By categorical facilitation, we refer to an aspect of categorical perception, in which the linguistic distinction between categories affects color discrimination beyond the low-level, sensory sensitivity to color differences. According to this idea, discrimination performance for colors that cross a category border should be better than for colors that belong to the same category when controlling for low-level sensitivity. We controlled for sensitivity by using colors that were equally discriminable according to empirically measured discrimination thresholds. To test for categorical facilitation, we measured response times and error rates in a speeded discrimination task for suprathreshold stimuli. Robust categorical facilitation occurred for five out of six categories with a group of inexperienced observers, namely for pink, orange, yellow, green, and purple. Categorical facilitation was robust against individual variations of categories or the laterality of target presentation. However, contradictory effects occurred in the blue category, most probably reflecting the difficulty to control effects of sensory mechanisms at the green–blue boundary. Moreover, a group of observers who were highly familiar with the discrimination task did not show consistent categorical facilitation in the other five categories. This trained group had much faster response times than the inexperienced group without any speed–accuracy trade-off. Additional analyses suggest that categorical facilitation occurs when observers pay attention to the categorical distinction but not when they respond automatically based on sensory feed-forward information.

Introduction

Does the distinction in language between “blue” and “purple” make it easier to see differences between bluish and purplish than between several purplish color shades? Evidence for categorical facilitation would give a positive answer to this question and show that language interacts with color perception.

The present study investigates whether the linguistic distinction between color categories affects the discrimination of colors. *Color categories* are the ensembles of different color shades that are grouped through a common color name, such as “blue” or “purple.” The linguistic distinction between two color categories, such as the difference between the blue and the purple category, define a *categorical difference*. In contrast to categorical differences, the differences that may be perceived between different color shades, such as between two different shades of purple, may be called *perceptual differences*. *Color discrimination* consists of identifying or detecting perceptual differences between colors.

In the context of the present study, it is important to consider the difference between the sensory ability to discriminate and the actual performance in a discrimination task. The sensory ability to discriminate two colors depends on the size of the perceptual difference between these colors: Two very different colors are easier to discriminate than two similar colors. The sensory ability to discriminate two colors of a given perceptual difference will be called *discriminability*. Hence, the discriminability of a given pair of colors is determined by the size of the perceptual difference and the *sensitivity to color differences*. The sensitivity to color differences is the basic ability of the visual system to detect perceptual differences between colors and may be measured through discrimination thresholds. A previous, related study showed that the sensitivity to color differences is not related to linguistic color categories and merely reflects low-level sensory mechanisms of color vision (Witzel & Gegenfurtner, 2013).

In contrast to sensory abilities, the actual performance in a discrimination task may be shaped by factors beyond the purely sensory determinants of discriminability. Although it is clear that discrimina-

Citation: Witzel, C., & Gegenfurtner, K. R. (2015). Categorical facilitation with equally discriminable colors. *Journal of Vision*, *15*(8):22, 1–33, doi:10.1167/15.8.22.

doi: 10.1167/15.8.22 Received July 29, 2014; published June 30, 2015 ISSN 1534-7362 © 2015 ARVO

bility will shape discrimination performance, other factors, such as attention and cognition, may also affect the performance in a concrete task. The present study tests this idea with respect to the influence of linguistic factors on discrimination performance. In particular, we investigate whether the linguistic distinction between categories affects discrimination performance.

We propose the notion of *categorical facilitation* to describe the idea that linguistic color categories affect discrimination performance in a way that may not be predicted by discriminability. According to this idea, the categorical difference facilitates the discrimination of colors that cross the category border as compared to those that belong to the same category. To control for discriminability, we used colors that were *equally discriminable* in terms of empirical measurements of sensitivity. Discriminability was determined by measuring the magnitude of *just-noticeable differences* (JNDs) for stimuli of different hues. A JND is a discrimination threshold that corresponds to the smallest difference between two colors that an observer is just able to perceive. Performance was measured in terms of response times and error rates in a speeded discrimination task with stimuli being two JNDs apart. The present study completes the preliminary investigations of Witzel, Hansen, and Gegenfurtner (2009).

Relevance

The present study aims to clarify the relationship between color perception and language. There is an explanatory gap between the continuous, three-dimensional perception of color on the one hand and the linguistic categorization of color through color terms on the other hand. Because of this explanatory gap, color has become the prime example to investigate the relationship between perception and language (R. W. Brown & Lenneberg, 1954; Kay & Regier, 2006; Lupyan, 2012). Research on the relationship between perception and language involves questions about *linguistic relativity*, *linguistic determinism*, and the *Sapir-Whorf hypothesis*. These ideas suggest a strong impact of language on perception and thought (Gumperz & Levinson, 1996; Gellatly, 1995; Kay & Kempton, 1984; Gentner & Goldin-Meadow, 2003; Deutscher, 2011). Moreover, research on the relationship between color perception and categorization also contributes to the question of *cognitive penetrability*, which consists of the question of whether knowledge influences perception (Collins & Olson, 2014). Finally, the relationship between low-level sensory color processing and high-level color processing is also a central problem in the field of color research (Gegenfurtner & Kiper, 2003; Gegenfurtner, 2003; Valberg, 2001; Fairchild, 1998; De Valois & De Valois, 1993). Taken together, these topics have been of major concern for multiple disciplines, including psychology, neuroscience, vision science, philosophy, linguistics, cultural anthropology, computer science, and engineering.

Evidence for *categorical perception* would bridge the gap between color categorization and color perception by linking categorical to perceptual differences. According to the idea of categorical perception, color differences should be perceived as more pronounced for colors that belong to different categories than for colors from the same category (Harnad, 1987; Goldstone & Hendrickson, 2010). For example, colors on each side of the boundary between blue and purple should be perceived as more different than colors within purple. Hence, in the case of categorical perception, there should be patterns in the way observers perceptually distinguish colors, patterns that are specific to the categories (*categorical patterns*).

Categorical patterns have already been found in the 1970s and 1980s (Bornstein, Kessen, & Weiskopf, 1976; Bornstein & Korda, 1984; Kay & Kempton, 1984). In particular, to establish a direct relationship between color perception and color categories, *category effects* have been investigated in tasks that involved color discrimination. According to the idea of a category effect, the presence of a category boundary between two colors should reinforce their difference and, hence, boost their discrimination. For example, two colors around the boundary between blue and purple should be discriminated faster and more reliably than two colors within the purple category. Therefore, response times and error rates should be lower for the discrimination of two colors on either side of a category boundary than for the discrimination of a comparable color pair within a category. Such categorical patterns were considered to be evidence for category effects on color perception.

Several studies provided evidence for such category effects on color discrimination (Bornstein & Korda, 1984; Daoutis, Pilling, & Davies, 2006; Holmes, Franklin, Clifford, & Davies, 2009; Winawer et al., 2007; Witthoft et al., 2003; Yokoi & Uchikawa, 2005; Yokoi, Nishimori, & Saida, 2008; Witzel & Gegenfurtner, 2011; Kay & Kempton, 1984). Moreover, recent studies found that category effects on speeded response times mainly occur on the right side of the visual field, presumably reflecting the hemispheric lateralization of language (e.g., Gilbert, Regier, Kay, & Ivry, 2006; Drivonikou et al., 2007; Franklin, Drivonikou, Bevis, et al., 2008; Franklin, Drivonikou, Clifford, et al., 2008; Roberson, Pak, & Hanley, 2008; Roberson & Pak, 2009; Zhou et al., 2010; Paluy, Gilbert, Baldo, Dronkers, & Ivry, 2011). Most of those studies concentrated on the green–blue boundary, and many used similar sets of colors based on Munsell chips. However, some other studies that did not use those Munsell chips could not find category effects on discrimination (A. M. Brown, Lindsey, & Guckes, 2011; Lindsey et al., 2010).

In order to show genuine category effects on color discrimination, it is crucial that the observed patterns in discrimination are specific to the categories. The ability to make a perceptual distinction between two colors—their discriminability—directly depends on the difference between the colors (Cavonius & Mollon, 1984; Mollon & Cavonius, 1986; Nagy & Sanchez, 1990; Rosenholtz, Nagy, & Bell, 2004; Bonnardel, van Leeuwen, & Flintham, 2007). The larger the difference between colors, the higher their discriminability and, hence, the better their discrimination. For this reason, the control of color differences is crucial for the investigation of categorical perception (Lucy & Shweder, 1979; see also experiment 2b in Roberson, Davies, & Davidoff, 2000; Witzel & Gegenfurtner, 2011). For this reason, previous studies controlled color differences between the colors of each color pair through

- Differences in wavelength (Bornstein et al., 1976) or dominant wavelengths (e.g., Bornstein & Korda, 1984)
- Ordinal steps in the Munsell color system (Bornstein & Korda, 1984; Kay & Kempton, 1984; Rosch Heider & Olivier, 1972; Davidoff, Davies, & Roberson, 1999; Roberson & Davidoff, 2000; Roberson et al., 2000; Roberson, Davidoff, Davies, & Shapiro, 2005; Roberson et al., 2008; Özgen & Davies, 2002; Pilling, Wiggett, Özgen, & Davies, 2003; Witthoft et al., 2003; Gilbert et al., 2006; Drivonikou et al., 2007; Franklin, Drivonikou, Bevis, et al., 2008; Franklin, Drivonikou, Clifford, et al., 2008; Yokoi et al., 2008; Holmes et al., 2009; Davidoff, Goldstein, Tharp, Wakui, & Fagot, 2012; Zhou et al., 2010; Paluy et al., 2011)
- The Uniform Color Scales of the Optical Society of America (OSA) (e.g., Yokoi & Uchikawa, 2005)
- Euclidean distances in CIELUV (ΔE_{Luv}) (e.g., Laws, Davies, & Andrews, 1995; Pilling et al., 2003; Yokoi & Uchikawa, 2005; Daoutis et al., 2006; Drivonikou et al., 2007; Roberson et al., 2005; Roberson, Hanley, & Pak, 2009)
- Euclidean distances in CIELAB space (ΔE_{Lab}) (e.g., A. M. Brown et al., 2011; Lindsey et al., 2010)

However, in order to draw any conclusion about the relationship between perception and categorization, differences between colors must be controlled in a perceptually meaningful way. Tests for category effects require a comparison between a measure of color differences that is bare of category effects and a measure of color discrimination that shows category effects.

The problem with all the approaches used in previous studies is that it is unclear how those measures of color differences relate to color perception in the first place. Wavelength differences are not indicative of perceived color differences. The Munsell and OSA system, CIELUV and CIELAB space coarsely approximate discriminability across color space. They cannot guarantee the equality of fine-grained perceptual differences. Moreover, these color systems and spaces are conceived to account for color appearance phenomena. As a result, they are prone to mix effects of categorical perception into the measure of discriminability (for details, see Witzel & Gegenfurtner, 2011, 2013). Finally, some of the most influential studies even neglected the issue of perceptual equidistance (e.g., Winawer et al., 2007) or failed to control color rendering (e.g., Gilbert et al., 2006; Drivonikou et al., 2007; Paluy et al., 2011). Consequently, those previous studies did not control perceptual differences in a meaningful way.

Several of those studies tried to circumvent the problem of controlling perceptual differences altogether. In these studies, identical stimuli were used in different experimental conditions, and the category effects were modulated by brain hemisphere, secondary task, specific language, etc. Although this is, in principle, quite elegant, in many cases there are serious flaws, or the results could not be replicated (see, for example, A. M. Brown et al., 2011; Witzel & Gegenfurtner, 2011).

Depending on which perceptual measures are used to control color differences, category effects may be investigated for different kinds of perceptual information (Witzel & Gegenfurtner, 2014). One kind of perceptual information is the discriminability of colors that results from the sensitivity to color differences. In a previous study (Witzel & Gegenfurtner, 2013), we have shown that the sensitivity to color differences does not follow a categorical pattern when color differences were determined according to low-level sensory information about color differences. Results suggested that sensitivity to color differences is purely perceptual in the sense that it is mainly determined through low-level early visual mechanisms (the so-called second-stage mechanisms) and does not imply effects of linguistic color categories (see also Bachy, Dias, Alleysson, & Bonnardel, 2012; Cropper, Kvansakul, & Little, 2013).

Consequently, if there are any category effects on color discrimination, they must occur in addition to the effects of sensitivity on discrimination. Such category effects may emerge if the linguistic distinction between categories interacts or combines with the perceptual distinction between different color shades (Bornstein & Korda, 1984). The interference between perceptual and categorical information may happen at higher, more cognitive levels of color perception that are beyond the low-level, early visual stages that determine color sensitivity (Roberson et al., 2009; Witzel & Gegenfurtner, 2013). In particular, color categories may influence discrimination beyond sensitivity if observers direct their attention toward categorical differences when doing a discrimination task (Gellatly, 1995; Deutscher, 2011). To distinguish these kinds of category effects from other category effects on color discrimination (Witzel & Gegenfurtner, 2014), such as effects on color sensitivity (Witzel & Gegenfurtner, 2013) or on subjective color appearance (Witzel & Gegenfurtner, 2012b), we call them *categorical facilitation effects*.

To investigate these kinds of high-level categorical facilitation effects, it is necessary to control for the low-level effects of sensitivity so as to disentangle sensory and categorical determinants of discrimination. None of the previous studies controlled appropriately for sensitivity. In fact, there is even evidence that the aforementioned green–blue Munsell chips, which were used as a prime example in previous studies, are biased toward spurious category effects (Witzel & Gegenfurtner, 2011). Moreover, all the previous studies completely neglected individual differences in color naming and discrimination (Witzel & Gegenfurtner, 2013; A. M. Brown et al., 2011). Consequently, the problem of controlling the sensitivity to color differences has yet to be solved in order to convincingly prove categorical facilitation effects on color discrimination.

Objective

The present study tested for categorical facilitation effects while controlling for the impact of color sensitivity on discrimination performance. To control for sensitivity, we used empirically measured JNDs to make color pairs equally discriminable. The rationale behind this approach is that two colors that can just be discriminated are identical in discriminability to two other colors that are also just discriminable.

To create the equally discriminable color pairs, the measurements of color categories and JNDs from the aforementioned study on categorical sensitivity were used (Witzel & Gegenfurtner, 2013). In that study, the task to measure JNDs was designed to measure low-level sensitivity. For this purpose, the difference between colors converged toward the JNDs during the task, observers were not under time pressure, and JNDs for each test color were measured in a separate block. In this way, participants were led to maximally exploit low-level sensory information about color differences.

To test for categorical facilitation in the present study, we measured performance in discriminating the equally discriminable color pairs through a *speeded discrimination task*. To guarantee that JNDs were a valid measure of low-level discriminability in the speeded discrimination task, this task was largely the same as the discrimination task for the measurement of JNDs. However, the speeded discrimination task differed in three important characteristics from the JND measurements in order to allow for potential categorical facilitation effects beyond low-level discriminability. First, instructions and feedback presented the speeded discrimination task in a game-like way that encouraged participants to respond as quickly as possible. Second, the equally discriminable color pairs used in this task were designed so that differences between the colors of each pair were clearly visible, suprathreshold differences. Finally, all equally discriminable color pairs, which involved fundamentally different regions of color space (see Equally discriminable color pairs section), were presented interleaved and in random order across trials. These three features were meant to encourage observers to combine information about perceptual and categorical differences in order to maximize their performance under time pressure, to prevent them from concentrating on fine-grained perceptual differences that require maximally exploiting low-level sensitivity, and to prevent them from tuning in to particular hue differences during each block of the speeded discrimination task.

Performance in the speeded discrimination task was measured in terms of response times and error rates. If discrimination performance is fully determined by the sensitivity to color differences, all equally discriminable color pairs should yield the same response times and error rates in the speeded discrimination task. In contrast, if color categories affect discrimination performance beyond color sensitivity, speeded discrimination should be better at the category boundaries than within the categories. Hence, a systematic decrease of response times and error rates toward the category borders must be attributed to categorical facilitation because the ability to discriminate these color differences based on low-level early visual mechanisms was equal across all color pairs.

However, color space is inherently anisotropic, which means that color differences in one part of the color space may have different characteristics than comparable differences in another part of the color space (e.g., Wuerger, Maloney, & Krauskopf, 1995). For this reason, we tested the predictions of categorical facilitation for all adjacent categories along an isoluminant hue circle. Keeping our color stimuli isoluminant controls for particular patterns in discrimination performance that may result from luminance changes due to different contributions of the achromatic (i.e., L+M) second-stage mechanism.

Moreover, the measurements were done with two groups of participants in order to account for possible effects of interindividual differences and experience with the discrimination task. The participants of the first group had previously participated in the extensive series of JND measurements, done in the study on categorical sensitivity (Witzel & Gegenfurtner, 2013). This allowed us to use individual measurements of categories and JNDs for the creation of equally discriminable color pairs so as to account for individual differences in sensitivity and categorization. At the same time, these observers were highly experienced with the discrimination task when speeded discrimination performance was measured with the equally discriminable colors. In this way, the measurements with this first group differed from previous studies on categorical perception, in which the same stimulus set was used for all observers and observers were mostly inexperienced with the task. To create conditions that were comparable with the previous studies, we tested for categorical facilitation effects with a second group of new observers. These observers were inexperienced in so far as they had not participated in the preliminary measurement of discrimination thresholds. Aggregated discrimination thresholds and categories of the observers in the first group were used to produce the same set of equally discriminable colors for each observer in the second group.

First, JNDs for 10 participants were previously measured by Witzel and Gegenfurtner (2013) in order to produce equally discriminable color pairs. Second, the main experiment consisted of the speeded discrimination task with the equally discriminable stimuli and tested for categorical facilitation effects. For this purpose, speeded discrimination was measured repeatedly across several sessions for each of nine participants in the first and 12 participants in the second group. The resulting large amount of data allowed for examining how categorical facilitation depends on response time distributions, individual differences, time and training, and lateralization. Finally, to establish a relationship between categories and speeded discrimination, it is crucial that the categories and JNDs assumed for the creation of equally discriminable colors were valid for the speeded discrimination task. For these reasons, additional measurements and analyses were done to verify the validity of the categories and JNDs for the speeded discrimination task.

Method

Equally discriminable color pairs

Preliminary measurements of JNDs and categories

All details about the measurement of categories and JNDs in the previous study have been described by Witzel and Gegenfurtner (2013). In brief, 10 observers (two male, eight female) participated in that study. Low-level sensory information about color at the second stage of color processing was modeled through Derrington-Krauskopf-Lennie (DKL) color space (Derrington, Krauskopf, & Lennie, 1984; Krauskopf, Williams, & Heeley, 1982). Stimulus colors were sampled from an isoluminant, saturated hue circle in DKL space. Colors had a luminance of 28 cd/m^2, and their saturation was high and roughly equal. The color of the background was achromatic (x = 0.31, y = 0.35) and had the same luminance as the stimuli.

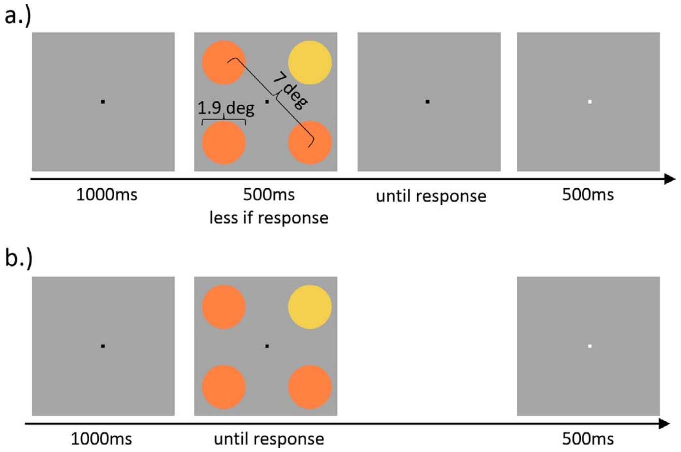

Figure 1. Discrimination task. (a) The time course of trials in the discrimination task for the measurements of JNDs and (b) the speeded discrimination task of the main experiment. Size and distances are indicated in the stimulus display of (a). They were the same in both tasks.

Color naming was measured for 120 colors along the hue circle with 3° azimuth between adjacent colors. Participants named the colors by using the eight chromatic basic color terms (pink, red, orange, yellow, green, blue, purple, and brown).

JNDs were measured for 72 equally spaced (5°) test colors along the hue circle. A four-alternative forced-choice (4AFC) discrimination procedure was used (cf. Krauskopf & Gegenfurtner, 1992). The task is illustrated by Figure 1a. Stimuli were rendered as colored disks of 1.9° visual angle. These disks were presented at four locations around a fixation point at the center of the screen. The distance between the centers of diagonal disks was 7°. Three of the disks, the distractors, were in the test color; the fourth disk was the target and had the comparison color. Participants had to indicate at which of the four positions the target was.

Each trial began with the presentation of a black fixation dot on the gray background for 1 s. Then, the stimulus display was presented for 500 ms or less if a response was given before 500 ms. If no response was given during the 500 ms, the display with the fixation point was shown until response. After the response, feedback about the correctness of the answer was given

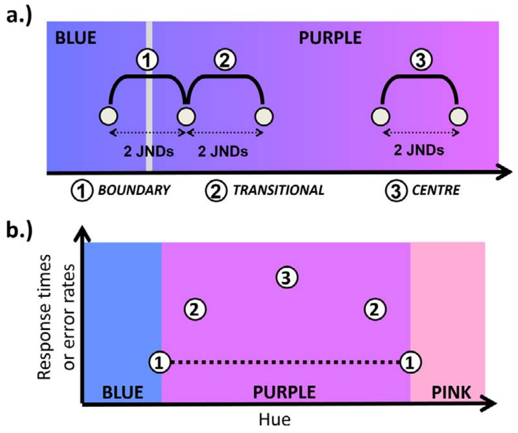

Figure 2. Design of equally discriminable color pairs. (a) The three kinds of equally discriminable color pairs at the blue–purple boundary. Gray disks correspond to particular colors; the black brackets illustrate the pairings. The study involved boundary pairs (1), transitional pairs (2), and center pairs (3). The two colors of all pairs differed by exactly two empirical JNDs (dotted arrows). (b) The predictions for those color pairs. Color pairs are shown for the complete purple category, ranging from blue (left boundary) to pink (right). As a result, there are two boundary (1) and two transitional pairs (2). The vertical axis corresponds to discrimination performance, measured in terms of response times and error rates. To account for global tendencies, response times are evaluated relative to the line that connects the response times at the boundary (*boundary line*). In the case of categorical facilitation, boundary pairs should yield lowest, center pairs highest, and transitional pairs intermediate response times. The same predictions apply to error rates.

by slightly changing the lightness of the fixation point for 500 ms.

Through a three-up-one-down staircase technique, this procedure converged toward a color difference that the observer detected with a probability of 0.79, which corresponds to a probability of 0.72 of seeing the difference (Levitt, 1971). To complete these measurements, each observer participated in 12 sessions of 45–60 min each. These measurements provide an exhaustive sample of JNDs in that average JNDs (about 8°) tended to be higher than the difference between test colors (5°). This allowed us to represent categories in JND space in which one unit corresponds to one JND and in which (small) distances are equally discriminable. We used this space to determine equally discriminable stimuli.

Design and predictions

Figure 2a illustrates the design of the equally discriminable color pairs. According to the idea of categorical perception, response times and error rates to discriminate between the two colors of a pair should be lower in pairs in which the two colors belong to different categories than in pairs in which the colors belong to the same category. To test this idea, we created three kinds of equally discriminable color pairs. First, a *boundary pair* consisted of two stimuli on one and the other side of a category boundary. This kind of stimulus pair corresponds to across- or between-category pairs in previous studies (e.g., Bornstein & Korda, 1984; Gilbert et al., 2006; see Introduction). These previous studies used a second stimulus pair, whose colors were on the same side of a boundary but shared one color with the boundary pair. These pairs were considered as within-category pairs because both colors were located in the same category. However, this kind of within-stimulus pair is still close to the boundary, and category membership is uncertain for colors close to the boundaries (Raskin, Maital, & Bornstein, 1983; Olkkonen, Witzel, Hansen, & Gegenfurtner, 2010; Witzel & Gegenfurtner, 2013; Huette & McMurray, 2010; Witzel & Gegenfurtner, 2011; Witzel, Hansen, & Gegenfurtner, 2008; Witzel, 2011). For this reason, we considered those category pairs as *transitional pairs* and added a third kind of stimulus pair. To obtain a stimulus pair that was unambiguously within a category, we created a *center pair*, whose colors were located around the center of a category.

In order to obtain clearly visible, suprathreshold color differences, we determined the two colors in each of these pairs so that they differed by exactly two empirical JNDs. The choice of a difference of two JNDs was a compromise between the requirement of suprathreshold stimuli and the nonlinearity of discriminability as a function of threshold differences. On the one hand, the nonlinearity of discriminability implies that the addition of JNDs has neither linearly scaled effects on the discrimination of the resulting colors, nor does it have the same effects all around the color space (e.g., Wuerger et al., 1995).

On the other hand, we needed suprathreshold differences to allow for categorical facilitation (see Introduction). When differences are at or below threshold, observers are not always able to detect the stimulus difference, and response times can become meaningless. This is illustrated by Supplementary Figure S1. It shows the response times of correct answers (triangles) during the measurements of JNDs at four test colors (black vertical lines). At one JND (dashed lines), the probability of seeing the difference is 0.72, and response times vary strongly and unsystematically. This shows that differences at threshold only allow measuring the probability of detecting the difference, and hence, they only allow for measuring the sensitivity to color differences as done in Witzel and Gegenfurtner (2013). Consequently, differences above threshold are necessary to measure facilitation effects independently of sensitivity. Supplementary Figure S1 shows that response times converge to stimulus offset time (500 ms) beyond two JNDs (solid red and green lines in Supplementary Figure S1). This indicates that two JNDs is the minimum difference that allows for measuring sensible response times.

Even though discriminability does not change linearly, it changes smoothly and continuously across hues (cf. Figure 3a here and figure 9 in Witzel & Gegenfurtner, 2013). This implies that local changes may well be linearly approximated. As a result, the addition of two JNDs should barely produce distortions in discriminability across stimulus pairs. Moreover, a distance of two JNDs implies one JND at each side of the discrimination center of each pair. So, all colors had definitely the same distance of one

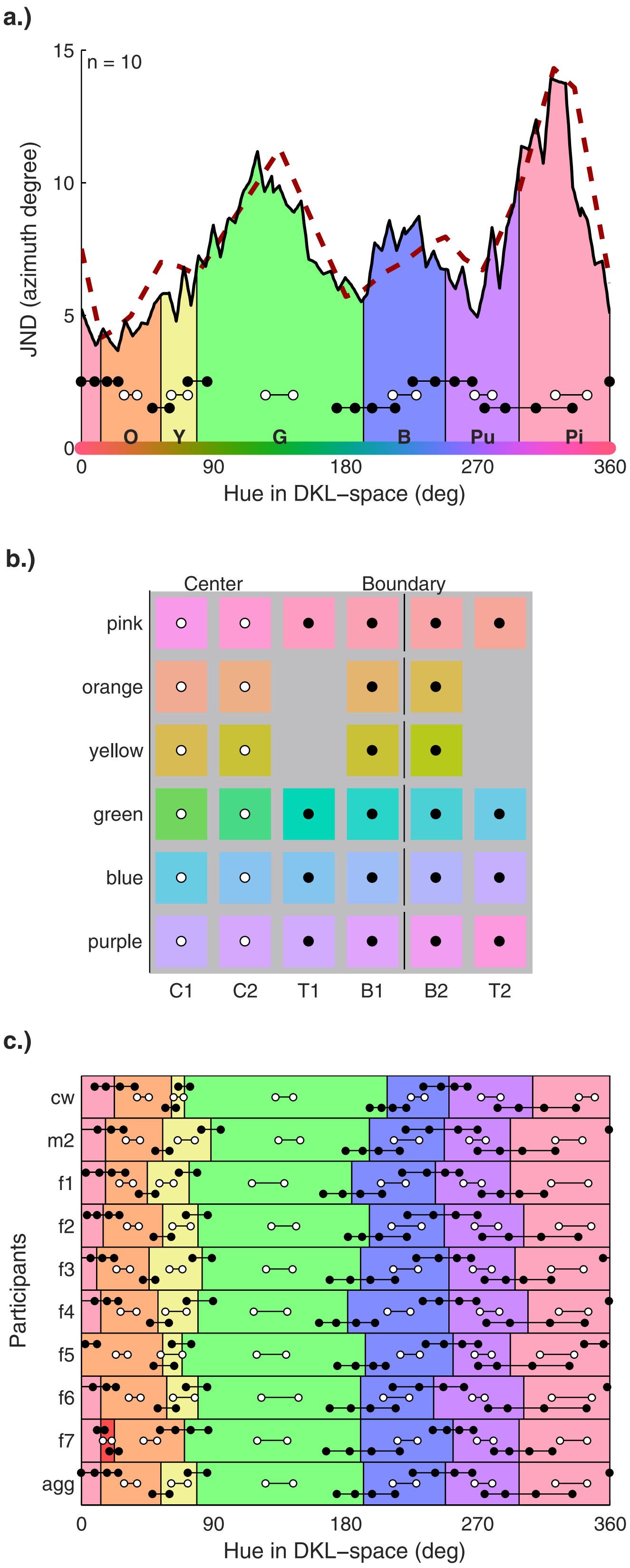

Figure 3. Results of preliminary measurements and production of equally discriminable color pairs. (a) The JND measurements and the resulting equally discriminable differences in DKL space. The x-axis represents the variation of hue along the isoluminant → ← hue circle in azimuth degree. The y-axis corresponds to JNDs of hue in azimuth degree. Colored areas and the vertical lines between them correspond to categories and their boundaries, respectively. The colors of the areas and the two uppercase initials at the bottom of the areas identify the single categories (from left to right: O = orange, Y = yellow, G = green, B = blue, Pu = purple, and Pi = pink). The solid black curve above the colored areas shows the JNDs measured in the preliminary JND measurements with the first group of participants. The black and white disks represent the colors of the equally discriminable color pairs based on aggregated data and used for the second group of participants. Disks that belong to the same color pair are connected by a horizontal line. White disks identify colors of center pairs, black ones those of transitional and boundary pairs. The dashed red curve indicates the JNDs measured in the post hoc JND measurements (cf. Discussion) with the second group of participants. (b) The stimulus colors that correspond to the disks in (a). They are arranged so that columns indicate membership to different kinds of color pairs (C = center, T = transitional, B = boundary), and rows show category membership. Display colors are meant to give a coarse idea about the stimulus colors. (c) Differences of categories and stimulus pairs across individual observers. The x-axis represents hue in DKL azimuth degree as in (a); rows along the y-axis correspond to nine participants of the main experiment and the last row to the aggregated data ("agg"). The aggregated data corresponds to the categories and stimulus pairs shown in (a). Categories (colored areas) and equally discriminable stimulus pairs (disks) are illustrated as in panel a. For the method, note that in regions with high JNDs (e.g., green, pink), the distances between stimulus colors are larger in DKL space to make them equally discriminable and vice versa (a). Moreover, note the absence of transitional pairs for orange, yellow, and red (f7 in c) due to their small width (all panels), and note the differences of stimulus pairs across individual observers (c). For results and discussion, note that the post hoc JND measurements (dashed red curve in panel a) were overall similar to the preliminary JND measurements (black curve in panel a).

discrimination threshold toward this discrimination center.

Figure 2b illustrates the predictions for these color pairs. According to the categorical facilitation effect, boundary pairs are expected to result in the lowest response times and error rates as compared to transitional and center pairs. Moreover, if category membership decreases toward the category boundaries, categorical similarity should be highest at the center of the category. Hence, the strongest support for categorical facilitation would require that center pairs yield the highest response times and error rates and transitional pairs range between boundary and center pairs in terms of discrimination performance.

Production of individual stimulus sets

Figure 3a shows the average categories and JNDs resulting from the preliminary measurements. The color naming measurements resulted in six adjacent color categories. For all but one participant, these categories were orange, yellow, green, blue, purple, and pink. One participant (f7 in Figure 3c) also identified a red category, but in turn, she did not find a yellow category that was large enough to produce a stimulus set (for a discussion of the lack of a consistent red category, see Witzel & Gegenfurtner, 2013).

The production of the stimulus pairs is illustrated by the disks in Figure 3a. Boundary pairs (black disks) consisted of the colors at a distance of one JND to each side of a boundary (vertical black lines); center pairs (white disks) consisted of colors at a distance of one JND to each side of the category center. The category center was determined as the average azimuth of the respective two boundaries. Because there were six categories, we obtained six boundary and six center pairs with overall $2 \times 12 = 24$ stimulus colors.

Transitional pairs (black disks in quadruples) shared one color with boundary pairs. For each of the two colors of a boundary pair, the second color of the transitional pair was the color at a distance of two JNDs toward the center of the respective category (black disks at each side of a quadruple). In general, there were two transitional pairs at each boundary and in each category. However, some categories were too narrow to allow for meaningful transitional pairs. In these cases, we did not produce transitional color pairs at the respective two borders. As a result, there were only eight transitional pairs, namely two at each of four boundaries (four quadruples of black disks). Because each of these transitional pairs shares one stimulus with the respective boundary pair, the transitional pairs required only eight additional stimulus colors. In sum, there were overall 20 stimulus pairs (six center, six boundary, eight transitional) and 32 different stimulus colors (12 center, 12 boundary, and eight additional ones for the transitional pairs). A set of stimuli is illustrated in Figure 3b. Note that the colors in this graphic depend on the display of print used to show them, and hence they may differ from those presented on the calibrated monitor in the experiment.

Main experiment

One version of the main experiment was implemented with personalized stimulus pairs that were adapted to the individual categories and discrimination thresholds of each observer. This version of the experiment involved nine of the 10 observers who participated in the JND measurements of Witzel and Gegenfurtner (2013). A second version of the main experiment used aggregated stimulus pairs based on average categories and JNDs and involved 12 new observers.

Participants

The nine observers of the first group consisted of seven women and two men with an average age of 22 years (±4 years *SD*). These observers were highly trained in that they participated in the extensive measurements of their discrimination thresholds. One of the 10 observers who participated in the preliminary experiment did not participate in the speeded discrimination experiment because she left Gießen (f5 in Witzel & Gegenfurtner, 2013). For this reason, IDs for the nine remaining participants (e.g., Figure 3c) match those of Witzel and Gegenfurtner (2013) with the exception that the original f6–8 have been mapped to f5–7 due to the dropout of the original f5. One of the nine observers was the author CW; all other participants were naïve as to the purpose of the experiment. The participant CW only took part in two of the five sessions of the speeded discrimination task.

For the second group, 12 new, naïve observers (six women, age 26 ± 4 years) were recruited. One of the observers (m2) only participated in three sessions. All observers were native speakers of German only. Color deficiency was excluded by means of the Ishihara tables (Ishihara, 2004). All naïve participants were students at Gießen University and participated for 8€ an hour.

Apparatus

The setup was the same as in Witzel and Gegenfurtner (2013). In sum, stimuli were presented on an Iiyama MA203DT monitor driven by an eight-bit NVIDIA graphics card, with a spatial resolution of 1152×864 pixels, and a refresh rate of 75 Hz. Color rendering was calibrated and gamma-corrected. Experiments were written in MatLab (The MathWorks

Inc., 2007) with the Psychophysics toolbox extensions (Pelli, 1997; Brainard, 1997). Precise timing of stimulus presentation was achieved by converting presentation times into refresh rates and synchronizing stimulus presentation with the refreshment of the screen. Responses were recorded by an ActiveWire device to enhance the precision of response time measurements (ActiveWire Inc., 2003). For the naming pretest (see "Procedure"), the input device was a specially modified numeric keypad whose keys displayed the color names instead of numbers.

Stimuli

Figure 3c summarizes the different stimulus sets (black and white disks) for each participant. Supplementary Figure S2 complements these figures with further illustrations of individual differences.

For the participants in the first group, we had measured the individual JNDs and category borders in the preliminary measurements (see section Equally discriminable color pairs). For these participants, personalized stimulus sets were produced on the basis of the individual JNDs and category boundaries. As a result, the actual colors of the stimulus sets differed across participants (cf. Figure 3c) while being equally discriminable according to the individual JNDs (cf. equal distances of disks in Supplementary Figure S2b).

The aggregated stimuli used for the second group of participants correspond to those shown in the last row of Figure 3c and Supplementary Figure S2b. These stimuli were created by averaging the JNDs as well as the category boundaries across the 10 participants of Witzel and Gegenfurtner (2013). As a result, all 12 participants of the second group saw the same colors. The computed Judd-corrected chromaticity coordinates for the aggregated stimuli are provided in the section Colorimetric specifications of stimuli of the supplementary material.

As may be seen in Figure 3c and Supplementary Figure S2b, the yellow category of the respective eight observers was too narrow to produce transitional pairs at both boundaries (orange–yellow and yellow–green). This was also the case for the aggregated stimuli (last row of Figure 3c and Supplementary Figure S2b). For the one subject with the red but no yellow category (f7), the red category was too narrow to produce transitional pairs at the boundaries (pink–red and red–orange).

Procedure

Apart from the measurement of response times and error rates in a speeded discrimination task, an experimental session also included a control measurement of color naming (see below) and a second part, in which data for another study on subjective appearance was collected (Witzel & Gegenfurtner, 2012b). These measurements were repeated in five sessions on different days. In all tasks and sessions, stimuli were counterbalanced, and the order of trials was randomized.

At the beginning of each session, participants were first given an oral overview of the experiment. Then they were provided with more detailed, standardized instructions on the screen. The time for reading through the instructions also guaranteed that people adapted to the gray background of the screen. Then, the naming task and the discrimination task followed. This part of each session lasted overall about 25 min.

Figure 1b illustrates the speeded discrimination task. There were four differences between the speeded discrimination task and the discrimination task of the JND measurements:

- The stimulus display stayed until an answer was given (Figure 1b) instead of a limited presentation time of only 500 ms (Figure 1a). This was done to avoid distorting response times through the disappearance of the display.
- There was no staircase but a constant stimuli technique with the suprathreshold, equally discriminable stimulus pairs. Target and distractors in one trial corresponded each to one of the colors of the equally discriminable stimulus pairs.
- All different color pairs were presented interleaved in each block.
- In order to measure speeded responses, participants were encouraged to respond as quickly as possible without reducing accuracy. For this purpose, we implemented a feedback after each block and a hall of fame after each session. Both were based on scores that combined response times and error rates. See section Feedback and hall of fame of the supplementary material for further details.

Apart from these four differences, the stimulus display and the task were the same as in the JND measurements.

In each session, participants completed three blocks. At the beginning of each session, participants completed a practice block with 10 random trials. In each block, each color of the 20 color pairs was presented once as a target and once as a distractor and at all four target positions. As a result, there were overall 4 (positions) × 20 (pairs) × 2 (targets) = 160 trials per block, 480 per session (three blocks) and overall 2,400 (five sessions) per participant. This corresponds to 120 speeded discrimination data (response times and accuracy) per stimulus pair for each participant. Exceptions were the participant CW of the trained group with only two sessions (48 trials per stimulus pair) and participant m2 of the untrained group with only three sessions (72 trials per stimulus pair).

Color naming control

The additional naming task remeasured color categories for the exact stimulus sets of the speeded discrimination task. Categorization may be affected by the differences in stimulus sampling, in particular if observers name colors in contrast to the colors they saw in other trials (*range effects*). The color categories used for the creation of the equally discriminable stimulus pairs were measured with 120 colors along the isoluminant circle in DKL space. However, during the speeded discrimination task, participants saw the sample of 32 equally discriminable colors. These two stimulus sets sample hues from very different distributions because color categories have different widths in DKL space. For example, the sample of 120 colors contained comparatively many green colors because the green category is particularly large in DKL space (cf. Figure 3a). Consequently, it is possible that the category boundaries assumed for the creation of the equally discriminable stimuli do not correspond to those of the stimuli in the speeded discrimination task.

The naming control measurement was used to assess the discrepancy between assumed and actual categories of the equally discriminable stimuli in the speeded discrimination task. For this reason, the only difference between these and the preliminary measurements was the stimulus set. It only included the 32 colors of the equally discriminable stimulus set (instead of 120). Apart from that, the task was the same as the one used by Witzel and Gegenfurtner (2013) to determine the category boundaries for the stimulus production (cf. section "Preliminary measurements of JNDs and categories"). Each color was shown as a disk on the gray background. To assign a color to a category, participants pressed one of eight keys on the special input device. In one session, each color was presented once, resulting in 160 measurements across the five sessions.

Post hoc measurements

In the preliminary measurements, categories and JNDs differed across observers (cf. Figure 3c and Supplementary Figure S2; for details, see Witzel & Gegenfurtner, 2013, pp. 6–15). To produce equally discriminable stimuli for the second group, aggregated categories and JNDs of the first group were used. Post hoc measurements were conducted to verify whether these categories and JNDs were valid for the second group.

To obtain comparable data for the second group, their color categories and JNDs were measured with the same methods as those used with the first group in Witzel and Gegenfurtner (2013) (cf. section Equally discriminable color pairs). Six observers (two women, age 26.5 ± 3.1 years) of the second group took part in these measurements after completion of the main experiment. The measurements were done across six sessions.

For the comparison with the preliminary categories of the first group, the post hoc color naming test involved the same set of 120 colors along the hue circle with 3° azimuth between adjacent colors. In each session, color naming was measured once for each color, resulting in overall 720 measurements, six per color.

The post hoc JND measurements differed from the original measurements with the first group only by the set of test colors. JNDs were measured for the 20 equally discriminable color pairs used in the speeded discrimination task. For this purpose, the centers of each stimulus pair were used as test colors, and JNDs were measured to either hue direction of these test colors. Two sessions were needed for one measurement of all test colors because only 10 test colors could be measured in one session. The measurements were repeated three times per test color across the six sessions.

Results

The first section provides the main results on categorical facilitation in the speeded discrimination task. The speeded discrimination task yielded qualitatively different results between the first and the second group of participants. To explore differences between the two groups and validate the main results, the second section provides additional analyses of (a) the overall performance, (b) individual data, (c) the distribution of response times, (d) the development of performance across time and task experience, and (e) the lateralization of category effects. In the third section, we verify the validity of the color categories assumed for the production of the stimulus pairs by examining the results from the control measurements of color categories. In the final section, we inspect the relationship between the JND measurements and the category effects in the speeded discrimination task to clarify where category effects come from.

Incorrect responses and response times above 2 s were discarded from the analyses of response times (this did not affect the main results). Although predicted category effects go in a particular direction, reported test statistics will be two-tailed in order to evaluate

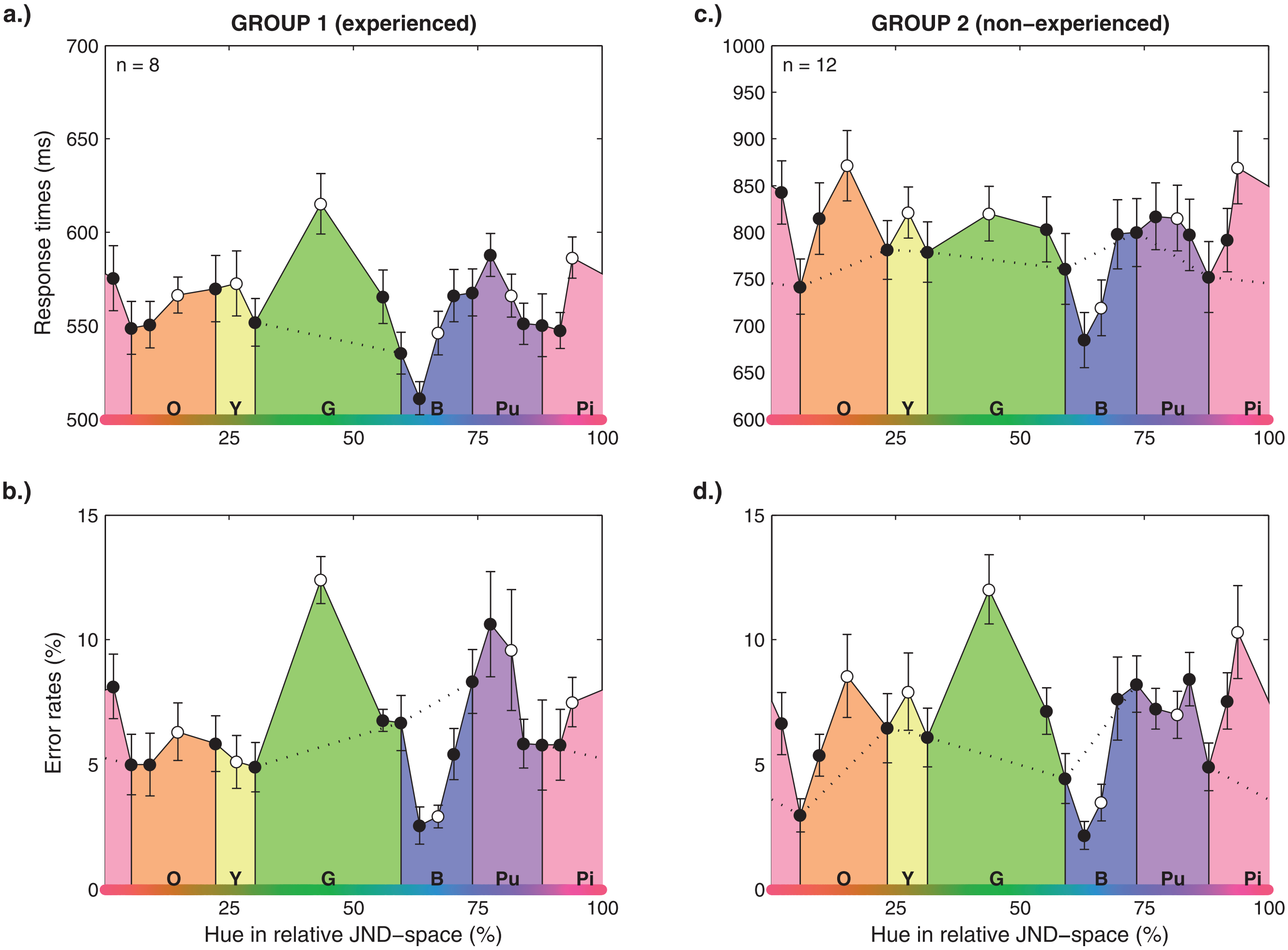

Figure 4. Average response times and error rates. The x-axis represents hue in relative JND space, specified in percentage of the overall hue circle. The y-axis corresponds to average response times (a and b) and error rates (c and d). Data was averaged across participants according to the stimulus pair type. White disks correspond to center, black disks to transitional and boundary pairs. Category boundaries are shown as vertical black lines, and category types are indicated by the color of the area under the curve and the initials of the color term (cf. Figure 3a). For illustration, boundary lines (the connection between the values at the boundaries) are shown as dotted lines when they do not cover the data. (a and b) The response times and error rates, respectively, for the eight participants with the same set of categories in the first group. (c and d) This data for the second group. Error bars correspond to standard errors of mean. In the second group, response times and error rates within the categories lay above the boundary lines for all categories except blue; the first group only shows this pattern for green and pink.

results in the direction opposite to the predicted category effects. In these two-tailed statistics, $p < 0.05$ will be considered as significant, $p < 0.01$ as highly significant, and $p < 0.1$ as marginally significant because it corresponds to $p < 0.05$ in a one-tailed test.

Main results: Category effects

Figure 4 illustrates the main results. To appreciate the distances between stimulus pairs in terms of discriminability, stimulus pairs are represented by their centroid in relative JND space along the x-axis. Relative JND space corresponds to cumulative JNDs starting from the azimuth of 0° relative to the overall number of JNDs along the whole hue circle (360°). To produce relative JND space, JND steps are divided by the total number of JNDs along the hue circle. In this way, JND differences are represented as proportions of the overall number of JNDs (see also section Individual differences in stimulus pairs of the supplementary material). To allow for comparisons, average response times (upper row) and error rates (lower row) are shown together for the first (left column) and the second group (right column). In the first group, only the eight participants with the same set of categories are shown. Data from the ninth participant did not differ in any systematic way.

	Response times (ms)				Error rates (%)			
	M	*df*	*t*	*p*	*M*	*df*	*t*	*p*
Pink	30.7	8	1.3	0.22	1.1	8	0.7	0.50
Orange	1.5	8	0.1	0.93	1.1	8	0.7	0.52
Yellow	12.4	7	0.5	0.60	0	7	0	0.99
Green	61.6	8	5.0	***	6.3	8	5.0	***
Blue	−3.0	8	−0.3	0.80	−4.2	8	−3.0	*
Purple	5.2	8	0.2	0.85	1.4	8	0.4	0.72

Table 1. Categorical perception tests for group 1. *Notes*: Paired, two-tailed *t* tests across observers testing whether relative response times (left part) and error rates (right party) of the center pairs were greater than zero. Symbols: M = mean, df = degrees of freedom, *t* = *t* value, *p* = chance probability; °, *, **, and *** correspond to $p < 0.1$, $p < 0.05$, $p < 0.01$, and $p < 0.001$. See Supplementary Table S2 for transitional pairs.

In both groups, response times and error rates were positively correlated across the 20 stimulus pairs (cf. upper and lower rows of Figure 4). To calculate the correlations between response times and error rates for the first group, response times and error rates were averaged across the eight participants with the same category set. The resulting correlation was $r(20) = 0.81$ and highly significant ($p < 0.001$). There was also a significant positive correlation for the ninth participant of the first group, $r(20) = 0.85$, $p < 0.001$. In the second group, the correlation between average response times and error rates was also highly significant, $r(20) = 0.80$, $p < 0.001$. These correlations indicate that performance varied systematically across stimulus pairs. They also show that there was no speed–accuracy trade-off. Hence, observers performed indeed better in discriminating some of the suprathreshold color pairs despite the fact that they were equalized in sensory discriminability.

According to the idea of categorical facilitation, response times and error rates should be lowest for boundary pairs and highest for center pairs. Hence, there should be a funnel-shaped pattern around the category boundaries (i.e., a categorical pattern). To test for such categorical patterns, response times and error rates were compared to the *boundary lines*. Boundary lines are the lines between the values at the boundaries (dotted black lines in Figure 4). They account for potential global modulations of performance that are not due to category effects. According to a category effect, the measurements within the categories are expected to lie above the boundary line. We tested for each category whether this was the case (*categorical perception tests* following Witzel & Gegenfurtner, 2013, pp. 16–17).

For this purpose, *relative* response times and error rates were determined as the difference of response times and error rates from the respective boundary lines. Paired, two-tailed *t* tests were used to test whether these relative response times and error rates were different from zero. To compare center and transitional pairs, *t* tests were applied to the difference between their relative response times and error rates, respectively. Tables 1 and 2 provide detailed results for the center pairs; those for the transitional pairs and the differences between center and transitional pairs are provided in the supplementary material (Supplementary Tables S2 and S3).

Categorical facilitation should affect all categories. Hence category effects were expected to appear for each category, not just for one of them. Observing the predicted categorical patterns in all six categories has a much lower probability than getting a pattern in just one category. For this reason, no correction for multiple testing across the six categories is applicable. Instead, if there were consistent categorical patterns in

	Response times (ms)				Error rates (%)			
	M	*df*	*t*	*p*	*M*	*df*	*t*	*p*
Pink	120.6	11	6.9	***	6.0	11	3.4	**
Orange	108.5	11	6.6	***	3.7	11	3.1	*
Yellow	41.3	11	4.5	***	1.7	11	2.2	°
Green	49.1	11	2.7	*	6.7	11	4.8	***
Blue	−61.3	11	−4.8	***	−2.9	11	−6.3	***
Purple	42.2	11	2.5	*	0.7	11	0.5	0.60

Table 2. Categorical perception tests for group 2 (inexperienced observers). *Notes*: Format as in Table 1. See Supplementary Table S3 for transitional pairs of the second group.

several categories, we tested whether the ensemble of categorical patterns was statistically significant. In some additional analyses beyond this main test for category effects, corrections for multiple testing were necessary, but this will be stated explicitly.

Group 1 (highly trained observers)

Table 1 and Supplementary Table S2 report the results of the categorical perception tests for the first group of observers, who were highly trained due to the preliminary JND measurements. Only in the green category, response times and error rates of the center pair were higher than the boundary line (p <0.001, cf. Table 1). In contrast, the pattern of the blue category contradicted a category effect (cf. Figure 4a and b). The transitional blue–green pair yielded the global minimum of response times and error rates. Response times, $t(8) = -3.4$, $p = 0.01$, and error rates, $t(8) = -2.3$, $p = 0.0496$, of this pair (cf. Supplementary Table S2) and the error rates of the blue center pair, $t(8) = 3.0$, $p = 0.02$, were significantly below the boundary line (cf. Table 1). These results for the first group do not support a consistent category effect across categories.

However, the results for the green category are still significant after multiplying probabilities by 12 as a Bonferroni correction for the 12 tests (response times and error rates for six categories). This indicates that the effect for green is significant even if there is no consistent evidence for category effects across all categories.

Group 2 (initially inexperienced observers)

Table 2 and Supplementary Table S3 provide the results of the categorical perception tests for the second group of observers, who were completely inexperienced with the discrimination task at the beginning of the main experiment. Pink, orange, yellow, green, and purple yielded the highest response times at the center pairs with the transitional pairs lying between the center and boundary pairs (cf. Figure 4c). The center pairs of these five categories yielded responses times above the boundary line (all ps < 0.04; cf. Table 2, left part). Category effects ranged between 41 ms and 121 ms. For the categories with transitional pairs (pink, orange, green, purple), response times of the transitional pairs also lay above the boundary line ($p < 0.05$; cf. Supplementary Table S3, left part). Relative response times for pink–purple and orange–pink were also significantly smaller than those for the respective center pairs ($p < 0.05$); other differences between transitional and center pairs did not reach significance (cf. Supplementary Table S3, left part). Hence, pink, orange, yellow, green, and purple show the funnel-shaped categorical pattern that is indicative for category effects.

Error rates mirrored the categorical pattern of response times for pink, orange, yellow, and green. Center and transitional pairs were also significantly above the boundary line (cf. Table 2 and Supplementary Table S3, right part) except for yellow, $t(11) = 2.2$, $p = 0.053$, and transitional orange–pink, $t(11) = 1.9$, $p = 0.08$, for which the difference was only marginally significant. For purple, only the transitional purple–pink pair yielded error rates above the boundary line, $t(11) = 2.5$, $p = 0.03$.

Results for blue yielded again a pattern that contradicted a category effect. The transitional blue–green pair corresponded to the global minimum. Both this transitional pair and the blue center pair resulted in response times and error rates below the boundary line ($p < 0.001$, cf. Table 2 and Supplementary Table S3). The transitional pair also yielded significantly lower relative response times and error rates than the center pair ($p < 0.05$, cf. Supplementary Table S3).

Because the pattern of results in the blue category contradicted a category effect, the question arises whether the categorical patterns in the other five categories may occur by chance, i.e., by random variation across observers. To approximate the chance probabilities of obtaining those categorical patterns, we used a binomial distribution with a chance probability of 0.05 (significance level) of obtaining a categorical pattern. Based on the binomial distribution, the probability for obtaining five out of six category effects with response times and four out of six with error rates are highly significant (both ps $\ll 0.0001$). When combining significant effects for response times (five) and error rates (four), the probability for nine out of 12 categorical patterns is still lower. These probabilities are significant even after multiplying them by two as a Bonferroni correction for the for the two groups of participants.

Clearly, these binomial statistics are merely approximations because we cannot guarantee that the patterns in the six categories are completely statistically independent (but see section Independence of categorical patterns of the supplementary material). However, in each category, it is not only the patterns of the center pairs that are in line with a category effect, but also the patterns of almost all transitional pairs (as in Supplementary Table S3). If we also take the patterns of the transitional pairs into consideration, the probability of getting the categorical patterns in Figure 4c and d is still lower than predicted based on the five patterns of the center pairs alone.

Apart from random variation across observers, there might also be random noise in the production of the aggregated stimulus pairs of the second group, for example, due to technical factors in stimulus rendering.

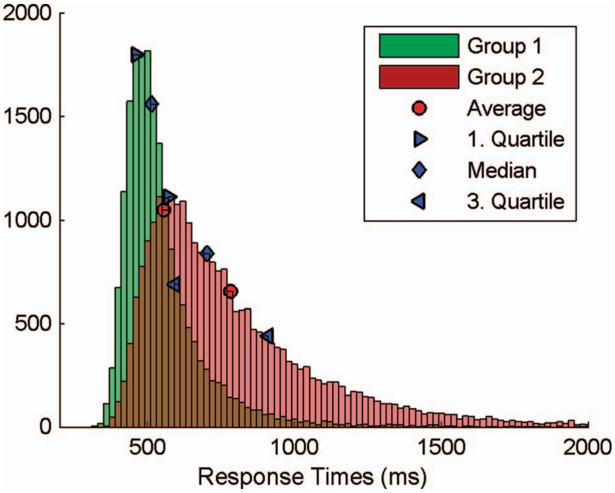

Figure 5. Histogram of response times. Response times of all participants of the first (green) and second group (red) are divided into 100 bins. The x-axis represents response times in milliseconds, the y-axis the frequency of the binned response times. The blue symbols show the quartiles, the red disks the average response time. Note that the main difference between the two groups consists of the higher amount of response times above 600 ms in the second than in the first group.

Such random effects across stimuli would result in systematic effects across individuals, and statistics to test such effects need to test for random variation across stimuli, not observers. Only the three stimuli in the blue category contradicted the categorical pattern of response times (Figure 4c). For error rates, the purple–blue transitional and the typical purple pair also contradicted a categorical pattern in the purple category (Figure 4d). We calculated the probabilities for obtaining at least 17 and at least 15 stimulus pairs out of 20 that are in line with a categorical pattern by chance based on a binomial distribution with a probability of 0.5 for a stimulus to be in line with the categorical pattern or not. These probabilities are both below the significance level of 0.05 ($p = 0.001$ and $p = 0.02$, respectively).

Hence, the categorical patterns found for the second group are unlikely to be due to unsystematic variations across either observers or stimulus pairs. These results confirm the presence of category effects in the performance of the second group. The following additional analyses further consolidate these observations.

Additional analyses of main results

The main results raise the questions of why consistent categorical patterns only occurred in the second but not in the first group and why the blue category in both groups yielded a pattern that contradicted the category effect.

Overall performance

To elucidate potential origins for the different category effects in the two groups, we compared the two groups in their overall performance. The average response times of the nine participants in the first group varied between 494 ms for the fastest and 605 ms for the slowest participants. The total average was 558 ms with a standard error of mean of 11 ms across participants. For the second group of initially inexperienced participants, the average response time across all participants and stimuli was 794 ms, ranging between 609 ms for the fastest and 1135 ms for the slowest participants with a standard error of 42 ms. The difference in response times between the two groups (255 ms) was significant in a two-sided t test comparing the averages across participants, $t(19) = 4.7$, $p < 0.001$. Both groups yielded similar error rates (6.3% and 6.7%, respectively) and low numbers of outliers (0.2% and 1.2%, respectively).

Figure 5 shows histograms (100 bins) of the response times (only corrects) for group 1 (green histogram) and group 2 (red histogram). In group 1, the median was 513 ms, the 75th percentile was at 592 ms, and most responses (1,560) were in the bin at 492 ms. In group 2, the median was 702 ms, the 75th percentile was at 910 ms, and most responses were in the bin at 552 ms. The two distributions mainly differed in the proportion of response times above 600 ms (cf. height of tails in both distributions). Together, these results show that the two groups have fundamentally different patterns of performance.

Individual observers

Differences in categorical facilitation between the two groups may be due to individual differences between the observers of the two groups. In this case, there should also be differences across observers within each group.

We provide thorough analyses of category effects at the individual level in the section Individual observers of the supplementary material. In sum, they show that consistent facilitation effects occurred for all observers in the second group (Supplementary Figure S4b). In the first group, only green and blue yielded consistent effects across observers: green in line with categorical facilitation, blue in the opposite direction. There were no individuals in the first group that showed consistent effects across categories (Supplementary Figure S4a).

These results further support the main results reported above. In particular, these individual analyses show that the absence of category effects in the first group is not due to individual differences and the smaller sample size of this group. Moreover, these findings further support the idea that there were fundamental differences between the two groups.

Response time distributions

Response times are not normally distributed, and the average response time might not be representative of the response time distribution. Moreover, a possible explanation for the variability in response time patterns is that category effects only occur at a certain response speed. In particular, the lack of consistent effects in the first group could be due to the quick response times of this group. For these reasons, we examined if categorical patterns emerge for slower response times in the first group and disappear for quick response times in the second group.

We provide detailed analyses of response time distributions and the relationship between response time distributions and category effects in the section Response time distribution of the supplementary material. In sum, results mainly confirm the main results shown in Figure 4. In the second group, there were categorical patterns for response times across the response time distribution (Supplementary Figures S6 and S7c and d), and in the first group, categorical patterns only occurred in the green category but in none of the others (Supplementary Figures S5 and S7a and b). Hence, the observed categorical patterns do not depend on the size of response times, indicating that category effects are robust across the response time distribution.

Time and training

An important difference between the two groups consisted in the familiarity with the discrimination task and the experimental setup. The participants of the first group were highly familiar with task and setup because they participated in 12 preliminary sessions for measuring their JNDs. In contrast, the second group was unfamiliar with the task when they came to the first session of the main experiment. Hence, differences in training and experience may have produced the differences in performance between the groups. In particular, we wondered whether category effects disappeared with increasing familiarity with the task. This would explain why there were only category effects for the inexperienced second but not for the highly trained first group. For this reason, we first inspected how performance in general changed over time in the two groups and second whether there were stronger category effects at the beginning of the measurements than at the end.

Detailed analyses are provided in the section Time and training of the supplementary material. In sum, the first group's performance followed an idiosyncratic pattern with a speed–accuracy trade-off over time (Supplementary Figure S8a and b). The second group improved in performance over time through an increase of speed at constant accuracy (Supplementary Figure S8c and d). Moreover, six participants in the second group and only two participants in the first group improved across blocks and sessions the scores used for the blockwise feedback and the hall of fame (Supplementary Figure S9). These results suggest that the second, inexperienced group, but not the first, experienced group, improved across blocks due to training and experience with the task.

However, there was no evidence for a modulation of category effects across the five sessions of the speeded discrimination task. The conditions (groups and categories) that yielded category effects did so across all sessions. Moreover, no additional category effects appeared for the first group when analyzing the first blocks of the speeded discrimination task (Supplementary Figure S10). If the difference in category effects between the two groups was due to effects of training and experience, then we should have found a modulation of category effects across time and training, but this was not the case. Hence, these results undermine the idea that categorical facilitation is affected by training and experience with the task.

Lateralization

According to the lateralized category effect, the category effect should appear exclusively or at least more strongly in the right visual field and not or less in the left. Lateralization effects would explain why categorical facilitation effects weaken when lumping together the data for the left and right visual fields. If there were strong lateralization effects in the first group, the absence of categorical facilitation on the left side could have covered the presence of such effects on the right side. Hence, the absence of significant category effects in the first group might be the result of combining the patterns of both visual fields. To test this idea, we examined whether the categorical facilitation effects studied here were lateralized. Detailed analyses are provided in the section Lateralization of the supplementary material.

There was some support for lateralization effects in the first group. For green, the categorical pattern was more pronounced on the right than on the left side. For orange, pink, and purple some tendencies toward a categorical pattern were found on the right side but none on the left (Supplementary Figures S11 and S13a and b).

However, apart from the green category, the lateralization effects in the first group contrasted the observation that this group did not yield reliable category effects in the first place. If the observed lateralization patterns were traces of genuine lateralized category effects, the second group should yield even stronger lateralization effects because they showed very pronounced category effects. But this was not the case (Supplementary Figures S12 and S13c and d). Moreover, blue yielded some patterns of lateralization (Supplementary Figure S13a and d). However, these patterns in the blue category completely contradicted any category effect. Finally, there was also no evidence that lateralized category effects were modulated over time (Supplementary Figure S14).

In sum, there were no systematic lateralization effects. Category effects of the second group were not lateralized, and lateralization effects of the first group occurred without category effects. These results suggest that lateralization effects are not linked to category effects. Given the multiple tests for lateralization, the occurrence of inconsistent lateralization effects in the data of the first group may be the result of random variation. For these reasons, the present results contradict the idea that category effects are lateralized. In particular, potential lateralization effects were not strong enough to explain the absence of consistent categorical facilitation effects in the first group.

Validation of color categories

We examined whether the variation of color categories across observers, stimulus sets, and sessions could potentially modulate category effects. The naming tests of the main experiment and the post hoc naming test allowed for assessing variations of categories and the impact of these variations on category effects.

Naming test of the main experiment

The naming test of the main experiment measured categories for the actual range of colors used in the speeded discrimination task. The categories assumed for the creation of equally discriminable color pairs were measured with a set of 120 colors in the preliminary measurements of Witzel and Gegenfurtner (2013). In contrast, the naming test of the main experiment was measured for the individual stimulus sets of 32 colors used in the speeded discrimination task. The comparison between these two naming measurements assessed the discrepancies between assumed and actual categories. Detailed analyses are provided in the supplementary material (section Naming test of main experiment).

Figure 6 compares the category membership of the 32 stimulus colors of the speeded discrimination task between the two kinds of measurements. The colors of the disks indicate the assumed category membership based on the preliminary measurements. The colored areas represent the mode color names across the five sessions of the naming test of the main experiment.

In both groups of participants, the category boundaries of this naming test do not completely agree with the boundaries of the preliminary measurements (for details, see Supplementary Figures S15 and S16). Discrepancies between the original and the remeasured categories suggest that categories slightly differed between the different stimulus sets of the preliminary naming test and the naming test of the main experiment.

Moreover, interindividual differences in categorization are illustrated by the differences across rows in Figure 6. For the second group, additional differences between the two measurements of color categories may result from differences in color naming between the groups because the aggregated color categories of the first group were used for the creation of equally discriminable stimuli of the second group.

Post hoc naming test

The post hoc naming measurements allowed the assessment of how strongly the individual categories of the observers in the second group differed from the aggregated categories of the first group. Details about the analyses of the post hoc naming tests are provided in the supplementary material (section Post hoc naming test). These supplementary analyses confirm differences between the individual categories of the second and the aggregated categories of the first group (cf. Supplementary Figure S17). Because the second group with the aggregated categories yielded category effects, these results support the idea that category effects occur for aggregated categories even if they differ from those of the individual observers.

Recategorization of stimulus pairs

To further test this idea, we examined whether the results in the speeded discrimination task would have been different if other measurements of categories were used to classify the equally discriminable color pairs. Details are provided in the supplementary material (section Recategorization of stimulus pairs). In sum, the results show that no new category effects appeared in the first group when recategorizing stimulus pairs either by aggregated categories (Supplementary Figure S18a) or by the categories measured through the naming test of the main experiment (Supplementary Figure S18b). In the second group, the same category

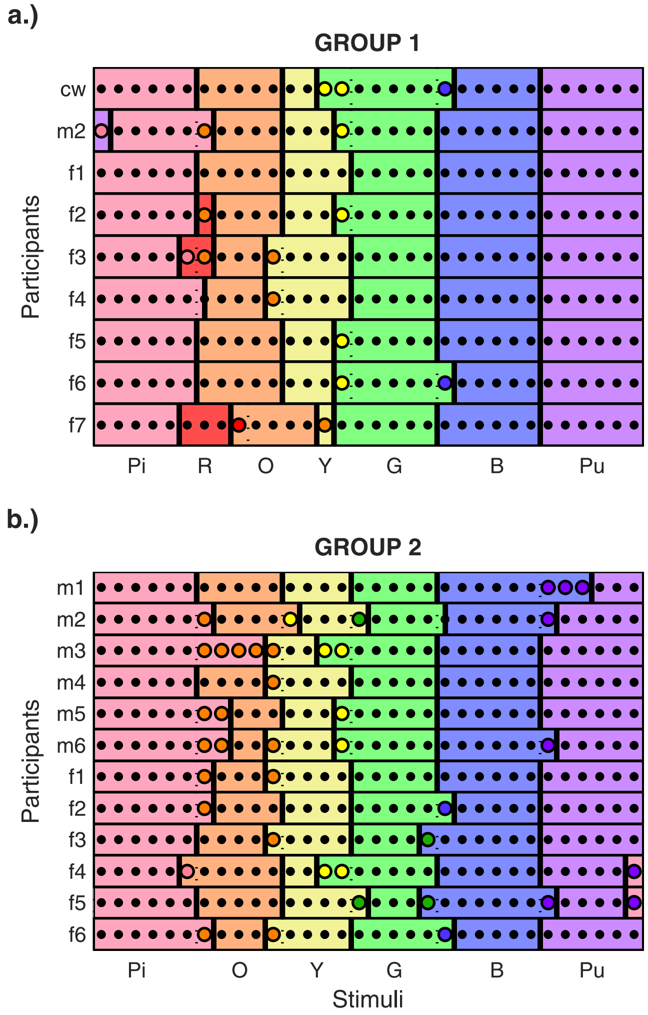

Figure 6. Categories in the speeded discrimination task. Graphics compare the categorization from the control naming tests (colored areas and thick black lines) to the category memberships of the equally discriminable colors (disks and dotted lines) that have been assumed based on the preliminary measurements of Witzel and Gegenfurtner (2013). Results for the first group are shown in (a) and those for the second group in (b). The x-axis corresponds to the 32 colors of each stimulus set, the y-axis to the individual observers of each group. Colored areas refer to the mode color names chosen by the respective observer. Thick black lines show the boundaries of the new categories, thin dotted lines those obtained in the preliminary measurements. In the first group (a), the dotted lines correspond to the boundaries of each individual's categories, in the second group (b) to the aggregated categories of the first group. The small black and the large colored disks refer to the equally discriminable stimuli. The large colored disks highlight those stimuli that yielded different color terms in the preliminary and the control naming task. Their color indicates the category membership measured previously and assumed for the stimulus creation. If newly measured categories were the same as the previously measured ones, the thick black lines would cover the dotted lines and comprise only black disks. However, categories slightly deviate from the assumed categories in both groups.

effects occur when classifying the stimuli by the individual categories of the naming test of the main experiment (Supplementary Figure S18c). Finally, in both groups, the blue category yielded the pattern that contradicts a category effect regardless of which categories were used to characterize the color pairs.

Taken together, these results show that the variation of color categories across observers, stimulus sets, and sessions is too small to affect category effects.

JNDs and speeded discrimination

In the present study, JNDs were assumed to reflect (low-level) discriminability and sensitivity to color differences while being not, or at least minimally, affected by categorical facilitation. The preliminary JNDs measured with the first group by Witzel and Gegenfurtner (2013) were used to define stimulus sets of both the first and the second group of participants. We wanted to verify whether the preliminary JND measurements provided reliable JNDs for the control of discriminability. Moreover, the differences between the two groups undermine the idea that categorical facilitation effects solely depend on the difference between measurements of JNDs and measurements of speeded discrimination as we had assumed. Hence, we wondered whether there were traces of category effects in the JND measurements. Finally, the development of the performance in the speeded discrimination task across time and training also disagreed with the idea that categorical facilitation simply declines with training and task experience. The question arises of whether JND measurements have particular training effects on discrimination performance that affect the performance in the speeded discrimination task.

For these reasons, we first compared the preliminary measurements to the post hoc JND measurements. Second, we inspected response times during JND measurements. Finally, we examined the development of discrimination performance (response times, error rates, and JNDs) across different blocks of JND measurements.

Just-Noticable Differences

The post hoc JND measurements allowed the verification of whether the JNDs of the first group were valid to control perceptual discriminability and the sensitivity to color differences. JNDs of the post hoc measurements for the six participants of the second group are illustrated by the dark red curve in Figure 3a. The curve looks very similar to the one measured by Witzel and Gegenfurtner (2013) with the 10 participants of the first group (solid black line in Figure 3). Both curves correlated strongly and positively across the 20 hues of the post hoc measurements, $r(20) = 0.91$, $p < 0.001$, $R^2 = 83\%$. This observation generally validates the JNDs measured with the first group. Consequently, post hoc JNDs show as few categorical patterns in DKL space as preliminary JNDs, namely only for the pink and green categories but not for orange, yellow, blue, and purple (cf. Witzel & Gegenfurtner, 2013).

In the section JNDs and speeded discrimination of the supplementary material, we thoroughly analyzed the relationship between the variation of JNDs and the variation of performance in the speeded discrimination task. Results showed that the patterns of JNDs across hues differed from the patterns of response times and error rates in the speeded discrimination task. These results support the idea that the control of discriminability through JNDs disentangled the performance in the speeded discrimination task from the sensitivity to color differences as assumed for testing categorical facilitation.

However, the preliminary and the post hoc JNDs were not completely the same. We assessed the potential impact of these differences on the control of discriminability in the speeded discrimination task. Detailed results are provided in the section Differences between preliminary and post hoc JNDs of the supplementary material. They show that the difference between preliminary and post hoc JND measurements were related to the performance of the second group in the speeded discrimination task (Supplementary Figure S19). These results are intriguing because they suggest that the JNDs of the first group did not allow for completely controlling discriminability in the speeded discrimination task for the second group. At the same time, they also imply that there were patterns in the post hoc JNDs of the second group that were specific to the categories. These observations suggest that the second group yielded JNDs with slightly stronger categorical patterns than the first group.

Response times in JND measurements

The faint categorical patterns in the contrast between preliminary and post hoc JND measurements (Supplementary Figure S19a) suggest that there might be traces of category effects during the post hoc JND measurements. These categorical patterns might not be visible in the red curve in Figure 3a because they are covered by the overall pattern of JNDs in DKL space. However, some of the trials in the JND measurements involved suprathreshold color differences, such as those used in the speeded discrimination task. Hence, the question arises whether categorical patterns, such as those found for the second group in the speeded discrimination task, also occurred during JND measurements for responses to suprathreshold differences.

To test this idea, we analyzed response times for suprathreshold color differences (i.e., color differences greater than one JND) in the JND measurements (cf. Supplementary Figure S1). Detailed results are provided in the section Response times in JND measurements of the supplementary material. Results show that the suprathreshold response times in the JND measurements showed a similar pattern across test colors as the JNDs measured in this task but not as the suprathreshold response times and error rates measured in the speeded discrimination task (Supplementary Figure S20). Consequently, those suprathreshold response times of the JND measurements show as few categorical patterns as the JNDs. This implies that JND measurements never showed category effects: not in the pattern of JNDs (see above) nor in the pattern of suprathreshold response times. Because the first group did not yield systematic category effects in the speeded discrimination task, the present results also imply that the first group did not yield category effects in any task. The fact that the contrast between the JNDs of the two groups showed some categorical patterns (Supplementary Figure S19a) suggests that categorical patterns are inherent to differences between the two groups.

The results of those analyses further elucidate the differences between the two groups (Supplementary Figure S20 and Figure 7). The first group was slightly slower during JND measurements than during the speeded discrimination task. In contrast, the second group was much slower in the speeded discrimination task than in the JND measurements. This latter result for the second group contrasts the fact that participants were explicitly encouraged to respond as fast as possible in the speeded discrimination task but not in the JND measurements.

Taken together, response times were comparatively fast in the JND measurements and discrimination task of the first group and in the JND measurements of the second group. These fast response times did not yield consistent category effects. At the same time, the second group's slow response times in the speeded discrimination task showed consistent category effects. Hence, the results suggest that the occurrence of category effects is neither specific to the group nor to the task. Instead, it might result from differences in how the two groups completed these tasks.

Development during JND measurements

In the second, but not in the first, group, the speeded discrimination task was done before the JND measurements. The fact that the JND measurements yielded lower response times than the speeded discrimination task in the second group raises the question of whether the task of the JND measurements has

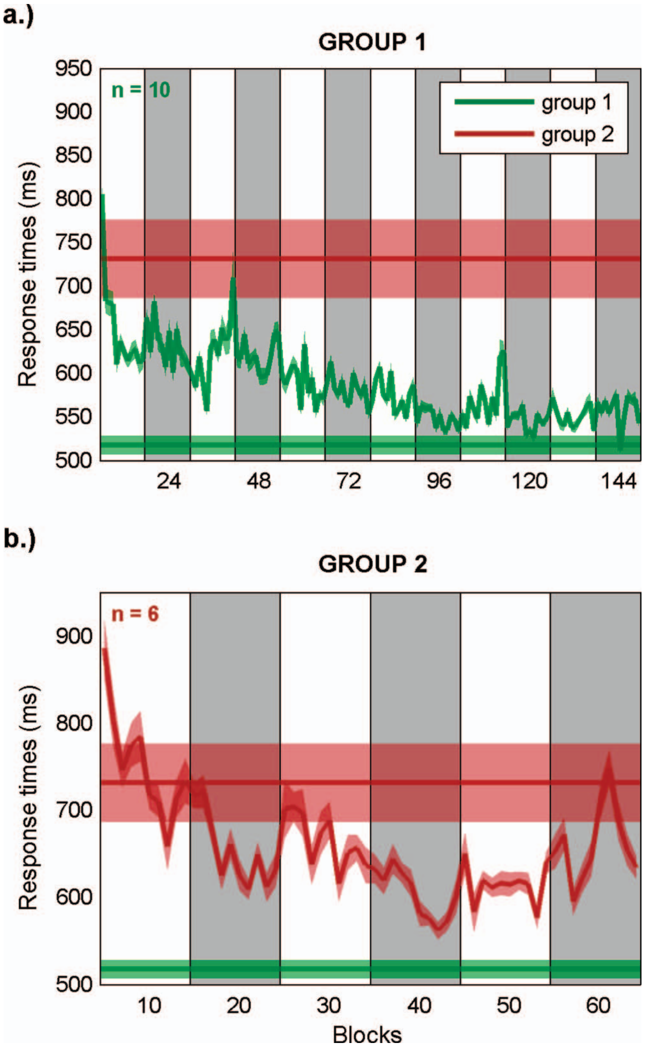

Figure 7. Development of response times during JND measurements. The x-axis corresponds to the blocks; the vertical white and gray areas in the background indicate the sessions. The y-axis refers to response times. Colored curves show the average response times for suprathreshold (greater than one JND) stimuli in the JND measurements across blocks and sessions. The red curve (a) corresponds to response times in the preliminary JND measurements of Witzel and Gegenfurtner (2013), and the green curve (b) to those in the post hoc JND measurements with the six participants of the second group. For comparison, median response times in the speeded discrimination task are illustrated by the red horizontal band for the first group and the green band for the second group. The horizontal lines represent averages, and the transparent areas standard errors of the mean of the median represented standard error of mean. Note that the first group started the JND measurements (green curve in a) at the level of response times of the second group in the speeded discrimination task (horizontal red line) and finished at a level slightly above its own response times in the speeded discrimination task (horizontal green line). In contrast, the second group never achieved the speed of the first group in any task (red curve and line in b).

stronger learning effects on response speed than the speeded discrimination task. In particular, the stimulus presentation time in the JND measurements was limited to 500 ms. This limit of presentation time might have exerted a stronger effect on response speed than instructions and feedback in the speeded discrimination task. This idea is supported by participants' reports. Participants first experienced the stimulus presentation in the JND measurements as extremely fast but got used to it after several trials. Most importantly, if the learning effects of the JND measurements counteracted category effects in a way that affects succeeding measurements, this could explain why the first group did not show any category effects in the speeded discrimination task.

To clarify the effects of the JND measurements on response speed, we examined how response times for suprathreshold (greater than one JND) color differences developed during the JND measurements. For this purpose, response times were aggregated by medians for each participant and block. Moreover, we compared the development of response times during the JND measurements to the response times of the speeded discrimination task in order to evaluate whether the experience with the JND measurements had particular effects on the performance in the speeded discrimination task.

The curves in Figure 7 show the the suprathreshold response times across the blocks of the preliminary (panel a) and the post hoc JND measurements (panel b). The green and the red horizontal bands refer to the response times of the first (green) and second (red) group in the speeded discrimination task. Detailed analyses are provided in the section Development of response times across blocks of the supplementary material.

Taken together, suprathreshold response times of both groups strongly diminish during JND measurements. The first group's response times in the speeded discrimination task were still lower than the lowest suprathreshold response times in the JND measurements (horizontal green line in Figure 7a). They provided the lower boundary for the response times of both groups during the JND measurements (green line in Figure 7b). This shows that the high level of response speed acquired during the JND measurements carried over to the speeded discrimination task. Hence, the experience of the first group with the JND measurements strongly affected their performance in the succeeding speeded discrimination task. In contrast, the second group's response times in the speeded discrimination task are higher than the response times in the JND measurements (red line in Figure 7b). These results suggest that the second group never reached the speed of the first group in the speeded discrimination task because they did not have the training with JND measurements before completing the speeded discrimination task.

We conducted further analyses to clarify whether the learning effects of the JND measurements counteracted category effects. Details are provided in the section Category effects across sessions of the supplementary material. In sum, at no point in time were there consistent category effects for the first group: neither in the JND pattern (Supplementary Figure S21a and b), nor in the suprathreshold response times of the JND measurements (Supplementary Figure S22a and b). The second group did not produce consistent category effects at any time during the JND measurements either (Supplementary Figures S21c and d and S22c and d), but there were some faint traces of stronger category effects at the beginning as compared to the end of the post hoc JND measurements (Supplementary Figure S22d).

These results suggest that JND measurements counteracted category effects. The lack of consistent category effects in the speeded discrimination task of the first group indicates that the suppression of category effects due to the extended experience with the JND measurements carried over to the speeded discrimination task. In contrast, the post hoc JND measurements followed the speeded discrimination task in the case of the second group. This second group showed consistent category effects in the speeded discrimination task. Hence, residual traces of these category effects might be the cause for the faint categorical patterns in the suprathreshold response times at the beginning of the post hoc JND measurements. These residual traces of category effects seem to disappear in the course of the post hoc JND measurements.

Discussion

There were categorical patterns for five out of six categories in the second, initially inexperienced, group (cf. Figure 4c and d). These patterns were robust across individuals (Supplementary Figure S4c and d), response time distributions (Supplementary Figures S6 and S7c and d), time course (Supplementary Figure S10c and d), laterality (Supplementary Figure S12), and variations in the set of color categories (Supplementary Figure S18c). It is highly unlikely that these patterns occurred by chance. Hence, these patterns reflect genuine category effects in the second group.

However, both groups' color pairs in the blue category yielded a pattern of response times and error rates that contradicted category effects (Figure 4). Moreover, there was a discrepancy between the presence of consistent categorical facilitation effects in the second group and the lack of such effects in the first group. In the first group, only the green category yielded a pattern of performance that was unambiguously in line with a categorical facilitation effect (Figure 4a and b). Finally, we did not find lateralized category effects unlike previous studies had suggested. Taken together, these results raise important questions about the determinants of category effects.

Control of discriminability

Increasing the magnitude of the differences between the colors to be discriminated should lead to faster and more accurate responses. To control for these effects of perceptual differences in the speeded discrimination task, we used sets of equally discriminable colors, in which the two colors of each pair were two JNDs away from each other. If the performance in the speeded discrimination task was completely controlled in its discriminability, response times and error rates should be constant across all color pairs in as far as they are not affected by categorical facilitation. Categorical facilitation would modulate performance specifically at category centers and boundaries. Hence, discrimination performance should only vary depending on whether color pairs were rather at the category centers or at the boundaries.

Nevertheless, in both groups, response times and error rates also varied across color pairs in a way that cannot be attributed to categorical facilitation. In particular, the pattern of performance in the blue category contradicted any category effects. Moreover, center pairs did not yield similar performance for all categories as would be expected if the only variation of performance was due to category effects (white disks in Figure 4). For example, green center pairs yielded worse performance than other center pairs. These results suggest that color pairs were not completely controlled in discriminability.

There are two possible origins of residual variations in discriminability across our color pairs. First, the relationship between threshold and suprathreshold differences is inherently nonlinear (e.g., Wuerger et al., 1995; Legge & Foley, 1980; Wilson, 1980). The addition of two JNDs should barely be affected by these nonlinearities because generally transitions in discriminability are smooth, and hence, local transitions may be linearly approximated with little error. However, it is possible that there might be abrupt changes for some particular color differences, which would even affect local transitions of only two JNDs.

Second, the JND measurements of Witzel and Gegenfurtner (2013) might have misrepresented the variation of JNDs across hues, for example, due to imprecisions and measurement noise. The addition of two JNDs would have doubled errors in measurement. The resulting variation in discriminability across color pairs might have affected the performance in the speeded discrimination task. This idea is supported by the fact that the post hoc JNDs differed from the preliminary JNDs in a way that correlated with the performance in the speeded discrimination task (cf. Supplementary Figure S19).

Categorical facilitation

The question arises of whether a failure to fully control JNDs produced spurious category effects. In particular, the differences in category effects between the first and the second group might be explained by a failure to control discriminability in the second group with the preliminary JNDs measured for the first group. This account for spurious categorical patterns is supported by the observation that the JNDs of the two groups slightly differed in a way that is roughly in line with category effects (Supplementary Figure S19a).

One possibility is that spurious category effects would result from the systematic variation of sensitivity across hues being not fully compensated in our stimulus production. In particular, the second group might have consisted of observers whose JND patterns across hues were more pronounced than the JND patterns in the first group. However, the pattern of JNDs across hues contradicts category effects on the sensitivity to color differences in DKL space as shown by Witzel and Gegenfurtner (2013) and by our additional analyses (Figure 3a and Supplementary Figures S20 and S21a and c). Hence, categorical patterns cannot be due to residual patterns of JNDs simply because JNDs do not have those categorical patterns.

Alternatively, unsystematic variation in the measurements of JNDs might have produced spurious category effects. There are two possibilities. On the one hand, spurious categorical patterns could be due to unsystematic variation of JNDs across observers. On the other hand, there might be random noise in the production of the aggregated stimulus pairs. However, we have shown that it is extremely unlikely that random variation across observers or across stimuli accidentally yielded the patterns in the five categories that were in line with category effects (cf. results of binomial tests in section Main results: Category effects).

Taken together, these observations show that the patterns across response times and error rates are systematic and specific to the categories. These categorical patterns cannot be due to failures to control discriminability and reflect genuine category effects. This conclusion is further supported by a follow-up study that showed—with different color sampling and different participants—category effects for equally discriminable colors at the red–brown boundary (Witzel & Gegenfurtner, 2012a). In particular, in that study, sensitivity was controlled with JNDs measured with the same participants for whom categorical patterns were shown in the speeded discrimination task. Hence, the traces of categorical patterns in the differences between the JNDs of the two groups (Supplementary Figure S19) cannot be the source of the categorical patterns in the speeded discrimination task.

Instead, the relationship between JND differences and category effects indicates that faint traces of the category effects also pervade the JND measurements of the second group. This finding implies that residual category effects from the speeded discrimination task carried over to the succeeding post hoc JND measurements of the second group (Supplementary Figure S22d).

At the same time, post hoc JNDs did not exhibit any more categorical patterns than preliminary JNDs in DKL space (dashed red curve in Figure 3a). This shows that residual category effects of post hoc JNDs were negligible compared to the overall pattern of JNDs. Hence, those post hoc JND measurements reconfirm the observation that the sensitivity to color differences is not categorical (Witzel & Gegenfurtner, 2013; see also Bachy et al., 2012; Cropper et al., 2013). This observation implies that the observed category effects in the second group cannot be due to categorical patterns in the sensitivity to color differences. Instead, they must be due to cognitive and linguistic factors beyond the sensory factors that shape color sensitivity (Witzel & Gegenfurtner, 2013; Roberson et al., 2009). Consequently, these category effects reflect genuine categorical facilitation.

In particular, these categorical facilitation effects may be explained by the idea that naïve, inexperienced observers, such as the ones in our second group, automatically direct their attention to categorical differences. This conclusion is further supported by recent studies that investigated category effects on event-related potentials (ERPs) when observers learned novel category boundaries (Clifford et al., 2012) and when color pairs were equally discriminable following our approach (He, Witzel, Forder, Clifford, & Franklin, 2014; Forder, He, Witzel, & Franklin, 2014). Although those studies did not find category effects in earlier ERP components that reflected perceptual processing, they found category effects in later components that correspond to postperceptual processes, such as attention.

An effect of attention to the linguistic distinction between categories implies that observers pay attention to the category boundaries that are specific to their language. This may explain observations of categorical patterns that vary depending on language (e.g., Roberson et al., 2000; Winawer et al., 2007; Kay &

Kempton, 1984) or depending on category learning (e.g., Özgen & Davies, 2002). Moreover, this understanding of categorical facilitation elucidates the modulation of category effects through verbal interference (Roberson & Davidoff, 2000; Pilling et al., 2003; Witthoft et al., 2003; Gilbert et al., 2006; Yokoi et al., 2008). Verbal interference occupies the observers' attention and hence may hinder the observers' ability to pay attention to the linguistic distinction between color categories.

Finally, previous studies found that response times for discriminating colors in visual search could be explained by color-opponent, second-stage mechanisms rather than by categories (Lindsey et al., 2010; A. M. Brown et al., 2011). Our findings in support of categorical facilitation complement rather than contradict those findings. Those studies equated color pairs through Euclidean distances in CIELAB space. CIELAB space coarsely controls global variations of discriminability across color space, but it may miss the local variations in sensitivity around the second-stage mechanisms that were observed by Witzel and Gegenfurtner (2013; cf. in particular figure 14). For this reason, those studies above may have obtained consistent local effects of the second-stage mechanisms despite equal differences in CIELAB space. In a second approach, those studies also measured the differences of color pairs in subjective appearance, using maximum likelihood difference scaling. Subjective appearance may also be affected by color categories (Kay & Kempton, 1984) at least to a small extent (Witzel & Gegenfurtner, 2012b, 2014). In this case, equating color pairs in subjective appearance should counteract potential category effects.

For these reasons, discrimination performance in those studies (Lindsey et al., 2010; A. M. Brown et al., 2011) should not be shaped by high-level category effects but by the sensitivity to color differences, which is strongly related to low-level second-stage mechanisms (Witzel & Gegenfurtner, 2013; Krauskopf & Gegenfurtner, 1992). In contrast, in the present study, we counteracted local effects of the second-stage mechanisms by using JND measurements to control for changes in sensitivity across hues. This allowed us to reveal high-level effects of categorical facilitation beyond variation in low-level sensitivity.

Nevertheless, the present study failed to completely disentangle categorical facilitation from measures of low-level sensitivity as shown by Supplementary Figure S19. A better separation between the two measures is a challenge for future studies.

Blue–green

In contrast to the effects of categorical facilitation, the peculiar pattern in the blue category may be explained by failures to control discriminability through the preliminary JNDs. In both groups, the blue–green transitional pair yielded the global minimum of reaction times and error rates, reflecting a maximum performance for this color pair. The adjacent blue center pair also yielded higher performance than the boundary pairs of the blue category (cf. Figure 4). These patterns contradict both categorical facilitation and the control of discriminability.

According to the post hoc JNDs, the distance between the colors of the blue–green transitional and the blue center pair were indeed larger than the JND distances of other color pairs (Supplementary Figure S19a). Hence, the preliminary measurements might have underestimated the JNDs for the green–blue transitional and the blue center pair. As a result, these color pairs would be more discriminable in reality than predicted by the preliminary JNDs. The particularly high performance for these color pairs in the speeded discrimination task might be simply explained by the fact that these pairs were easier to discriminate than other color pairs. If this is true, the absence of category effects in the blue category is most likely due to failures to control the perceptual determinants of discriminability close to the green–blue boundary.

Alternatively, variation of chroma and saturation across hues is another factor that could have modulated the performance in the speeded discrimination task independently of category effects. A hue circle in DKL space does not control well for chroma and saturation across hues well. Variations in chroma and saturation imply differences in salience (for details, see, e.g., Witzel & Franklin, 2014). If salience was particularly high around the blue–green transitional pair, this could result in a higher response speed independent of discriminability and category effects. We are still investigating this idea in an ongoing study.

In any case, other studies have also had difficulties showing category effects at the green–blue boundary when controlling for perceptual differences in terms of discriminability. Like the present study, the ERP study of He et al. (2014) involved equally discriminable color pairs at the green–blue boundary. Although that study found evidence for category effects in ERPs, there was no evidence for category effects in discrimination performance. The study of A. M. Brown and colleagues (2011) did not find category effects at the green–blue boundary when equating color differences in CIELAB. Hence, the lack of behavioral category effects at the green–blue boundary seems not to be a particularity of our study. Instead, it might reflect a general difficulty to control discriminability in the green–blue region of color space.

This difficulty may be due to nonlinearities at the green–blue boundary. When measuring sensitivity in

DKL space, JNDs are particularly low at the green–blue boundary and abruptly increase toward the center of the green and the blue categories (Figure 3a). Unlike other category boundaries, this boundary coincides with a second-stage mechanism, the L-M mechanism. This mechanism might be the origin for the particular variation of sensitivity around the green–blue boundary (Witzel & Gegenfurtner, 2013; see also Lindsey et al., 2010; A. M. Brown et al., 2011).

The strong variation of sensitivity in the green–blue region may also involve stronger local variations and nonlinearities in the relationship between threshold and suprathreshold differences. As a result, the linear approximation of suprathreshold differences through the addition of two JNDs might be more error prone in this than in other regions of color space. In particular, the control of discriminability through JNDs might overcompensate for the high sensitivity at the green–blue boundary and the comparatively low sensitivity for adjacent colors in the blue category. In this case, the addition of two JNDs would underestimate the suprathreshold distance at the green–blue boundary and overestimate those in the blue category. Consequently, this approach to estimate discriminability would produce blue–green transitional and blue center pairs that are particularly easy to discriminate compared to the green–blue boundary pair. Such nonlinear effects could explain our results for both groups at the green–blue boundary.

Most of the previous studies on the categorical perception of color investigated the green–blue boundary, assuming that it is representative for any other category boundary (Bornstein & Korda, 1984; Kay & Kempton, 1984; Gilbert et al., 2006; Siok et al., 2009; Drivonikou et al., 2007; Franklin, Drivonikou, Bevis, et al., 2008; Franklin, Drivonikou, Clifford, et al., 2008; Roberson et al., 2009; Holmes et al., 2009; Fonteneau & Davidoff, 2007; Özgen & Davies, 2002). In contrast to our results, those studies observed patterns that were in line with category effects. Because those studies did not control for variations in sensitivity across color pairs, their results are perfectly in line with those found here. The lack of control for differences in sensitivity resulted in green–blue boundary pairs that were easier to discriminate than adjacent color pairs (Witzel & Gegenfurtner, 2011, 2013). Consequently, the patterns observed in those studies might have been due to the fact that their boundary pairs were easier to discriminate rather than to genuine category effects.

The comparison of the response times that yielded categorical facilitation in our study and the response times that yielded category-like patterns in those previous studies further supports the idea that some of the previous evidence for category effects was spurious. In our study, categorical facilitation occurred in the second group that was untrained and had comparatively high response times (>700 ms). Few patterns of categorical facilitation occurred in the first group that responded at about 500 ms.

In contrast to these results, previous studies found categorical patterns with fast rather than slow response times when using stimuli that were poorly controlled in perceptual distances (e.g., Drivonikou et al., 2007; Roberson et al., 2008; Roberson & Pak, 2009; Siok et al., 2009; Witzel & Gegenfurtner, 2011; Zhou et al., 2010). Some of the studies that used the green–blue Munsell chips, which are biased toward spurious category effects (for details, see Witzel & Gegenfurtner, 2011), obtained those patterns with response times that were even lower than those of our first group (Gilbert et al., 2006).

Effects of perceptual differences would affect response times of all sizes. Our results suggest that genuine category effects only occur for untrained, inexperienced observers with high response times. If this is true, the low response times in previous studies reflect the fact that observers used perceptual rather than categorical information and that the effects in those studies occurred because of differences in discriminability.

However, several of those studies showed interaction effects, in which category effects at the green–blue boundary occurred specifically in an experimental condition that allows for the influence of language but not in a condition that excludes the influence of language. Because both experimental conditions involved the same kind of stimuli, the effects in those studies cannot be simply explained by a failure to control for discriminability. At the same time, not all of those interaction effects are equally convincing as exemplified by the discussion of the lateralized category effect below (section Is there a lateralized category effect?). For these reasons, it cannot be confirmed with certainty where the effects in each of those previous studies come from.

In any case, our findings suggest that it is particularly difficult to control discriminability and to reveal genuine category effects at the green–blue boundary. This difficulty is most probably due to the coincidence of this boundary with the L-M mechanism. Hence, the assumption that discrimination performance at this boundary is representative for all category boundaries does not hold.

More generally, the results at the green–blue boundary also highlight the fact that the control of suprathreshold differences through JNDs must be understood as an approximation. At the same time, the method of equalizing color differences through empirical JNDs is certainly superior to the approaches used in previous studies, which used color order systems or color spaces that coarsely approximate empirical JNDs. Furthermore, these findings show that category effects must be measured for a wide range of colors and categories to account for systematic effects of the nonlinearities of discriminability in some regions of color space.

Individual differences

The two groups differed by the individual observers that constituted the groups. Because categorical facilitation is probably a cognitive and linguistic effect, individual differences might play a comparatively strong role for categorical facilitation. For example, individual observers may strongly differ in motivation, thinking, and prior experience, and this might influence the extent to which their performance relies on linguistic or purely perceptual performance. However, categorical facilitation occurred for all observers in the second group (Supplementary Figure S4c and d), and the strength of categorical patterns did not vary systematically across observers (Supplementary Table S4). Hence, differences in the characteristics of participants seem not to affect categorical facilitation. In particular, such differences cannot explain the discrepancy in categorical facilitation between the two groups (Figure 4).

Variability of categories

An important difference between the first and the second group is that the determination of the stimulus pairs was based on individual color categories in the first and on aggregated categories in the latter case (Figure 3c). In general, there are individual differences in color categorization as may be seen in Figures 3c and 6 (see also figure 11 in Olkkonen et al., 2010, and figures 6 and 7 in Witzel & Gegenfurtner, 2013). In particular, the naming test of the main experiment (Figure 6) and the post hoc naming (Supplementary Figure S17) in the present study indicated that the individual color categories of the second group differed from the aggregated ones of the first group. However, it is not clear whether individual or aggregated categories are more relevant for category effects (cf. Witzel & Gegenfurtner, 2013).

On the one hand, individual category borders may be more relevant for the individuals' perception of color. To elicit category effects, categorization and perception must interact in the individual observer. More precisely, this should happen in the individual's brain as has been shown for phonetic boundaries (Chang et al., 2010). According to this reasoning, the individual observer's interpretation of the color terms should be more important than the consensus categories that reflect the commonalities across individuals. Hence, category effects should rather occur for individual than for consensus categories. However, the contrary was the case. Category effects occurred in the second group with the consensus stimuli instead of the first group with the individual stimuli.

On the other hand, aggregated categories could be more valid to test category effects for two reasons. First, the averaged categories are less noisy than the individual ones; this is particularly true if we take into account that category membership at the category border is not sharp but follows a probability distribution (cf. Olkkonen et al., 2010; Witzel et al., 2008). Second, in communication, the most important is the interindividual consensus, not the individual idiosyncrasies. Idiosyncrasies of categorization might develop due to personal experiences, but it is the consensus across individuals that allows for communication. For this reason, the interindividual consensus should be more stable due to its communicative function. As a result, consensus categories might be more prone to produce categorical facilitation. In this case, it is possible that linguistic category effects arise in the second rather than the first group because the consensus categories are more representative for color language than the individual ones (cf. Witzel & Gegenfurtner, 2013).

However, the application of consensus categories did not reveal additional, or at least stronger, category effects in the first group (Supplementary Figure S18a), and categorical facilitation effects for all five categories still appeared in the response times and error rates when the data of the second group was recategorized by the individual categories of the naming test (Supplementary Figure S18c). Hence, the results on categorical facilitation in the two groups do not depend on the difference between individual and consensus categories.

In addition to individual differences in categorization, category membership of colors close to the boundary is sensitive to the context in which the colors are shown. Category membership may be influenced, for example, by illumination changes (Olkkonen et al., 2010), by variations of background colors and adaptation (Witzel & Gegenfurtner, 2013, 2011), or by the color-diagnostic objects on which they are shown (Mitterer & de Ruiter, 2008; Mitterer, Horschig, Musseler, & Majid, 2009). Most importantly, category boundaries change depending on the task and in particular on the range of colors used in the naming tasks (Witzel & Gegenfurtner, 2011). In the present study, preliminary categories for the production of equally discriminable color pairs were measured with a range of colors that differed from the actual stimulus set in the speeded discrimination task. The naming test of the main experiment confirmed that category boundaries might vary across stimulus sets.

However, the range of variability in categorization is small compared to the size of categories (Figure 3c;

Supplementary Figures S15 and S16). The category effects observed in the second group involved the whole categories in that response times and error rates tended to increase toward the centers of the categories (Figure 4). Hence, it is unlikely that small shifts of category borders would strongly interfere with effects of categorical facilitation. This idea is confirmed by the recategorization of the data (Supplementary Figure S18).

Taken together, these results suggest that categorical facilitation effects were robust to the small effects of stimulus sampling and that they appear for both individual and consensus categories. Consequently, the differences in categorical facilitation between the two groups of participants cannot be due to the differences between individual and consensus categories.

Variability in sensitivity

It is rather surprising that there were systematic differences between the JNDs of the two groups (as in Supplementary Figure S19a). Because the method of the JND measurements was exactly the same for both groups, differences in JNDs may not be due to differences in the tasks. Moreover, differences in sensitivity across observers are small compared to variations in sensitivity across colors (Witzel & Gegenfurtner, 2013). The global pattern of JNDs across hues is very similar across observers (cf. Supplementary Figure S2a, and figure 3 in Witzel & Gegenfurtner, 2013). Individual observers mainly differ in the overall size of their JNDs and the strength of the JND pattern across hues (figure 8 in Witzel & Gegenfurtner, 2013). Moreover, the average post hoc JNDs in the present study were very similar to the preliminary JNDs, showing that there were only small differences between the individuals of the two groups (Figure 3a).

Nevertheless, the small differences between post hoc and preliminary JNDs (Supplementary Figure S19a) might have been due to differences in sensitivity across the observers in the first and second groups. Trichromatic observers may vary in fundamental physiological characteristics of color vision, such as the optic of the eye, the relative proportions of photoreceptors, and even their spectral sensitivities (for review, see e.g., Neitz & Neitz, 2011; Witzel, 2011). As a result of these physiological variations, small differences in color sensitivity may occur across normal, trichromatic observers. However, differences between the two sets of measurements were not specific to the cone-opponent mechanisms (cf. hues at 0°, 90°, 180°, and 270° in Supplementary Figure S19a). For this reason, these differences between JND measurements seem not to be directly related to differences in the characteristics of the photoreceptors or the cone-opponent mechanisms. Instead the small differences between preliminary and post hoc JNDs were rather in line with category effects, indicating residual traces of categorical facilitation in the second in contrast to the first group (cf. section on Categorical facilitation).

Finally, given the high similarity across individuals in the pattern of JNDs, the use of aggregated instead of individual JNDs seems appropriate to provide an approximate control of discriminability. In particular, if individual data sets involve a limited amount of measurements, the use of aggregated data sets for the control of discriminability across colors might be more effective because of less measurement noise.

Training and task demands

The two groups differed in the amount of experience with the 4AFC discrimination task before they started the speeded discrimination measurements. Moreover, they differed in the sequence in which they completed the measurements of JNDs and speeded discrimination. The question arises whether and how these differences may have modulated categorical facilitation so as to produce different categorical facilitation effects in the two groups (Figure 4).

Training and experience with the task

In general, our results showed strong effects of training and experience on discrimination performance. Observers improved in performance with increasing familiarity with the 4AFC discrimination task in both the JND measurements (Figure 7) and in the measurements of speeded discrimination (Supplementary Figures S8 and S9). The fact that the first group did not improve in performance during the speeded discrimination task indicates that they reached a ceiling effect in performance due to the preliminary experience with the extensive measurement of JNDs across 12 sessions. The second group did not have this experience when completing the speeded discrimination task. The initially untrained, inexperienced second group, but not the trained first group, yielded category effects in the speeded discrimination task (Figure 4). Hence, the experience with the extensive JND measurements seems to have counteracted categorical facilitation.

One possible reason for an attenuating effect of training on categorical facilitation may be the compression of response time variability. In fact, the average response times per stimulus pair for the first group covered less than half the range of those for the second group (~100 ms vs. ~250 ms; cf. Figure 5). The compression of response times through learning might prevent categorical facilitation effects because they depend on response time differences across stimuli. However, the second group yielded categorical facilitation effects on response times and error rates at all deciles of the response time distribution even though response time variability is strongly compressed in lower deciles (Supplementary Figure S7c and d). Hence, the size and variability of response times may not be the main factor that modulates categorical facilitation.

Alternatively, perceptual learning might have counteracted categorical facilitation. In perceptual learning, observers improve in very basic discrimination tasks through repeated experience with that task (Fahle, 2005). In particular, extensive training and experience with the task may improve the observer's ability to separate the signal that is relevant for discrimination from perceptual noise (Heinrich, Kruger, & Bach, 2011; Z. L. Lu, Hua, Huang, Zhou, & Dosher, 2011). In this way, the observer obtains a more reliable perceptual signal through training and experience.

In our study, the observers may learn which color differences occur in the discrimination task. With increasing knowledge about the color differences, they may search for those perceptual differences. This implies that they pay increasing attention to the perceptual rather than to the categorical differences with increasing experience with the task.

As a result, the categorical distinction does not, or at least considerably less, interfere with perceptual discrimination in highly trained observers. In this way, categorical facilitation might have disappeared in the trained observers of the first group due to their experience with the extensive JND measurements.

However, training and experience with the task cannot be the only explanation for differences in category effects between the two groups. Training and experience did not always attenuate categorical facilitation as shown by the result that the category effects of the second group did not reduce across the five sessions of the speeded discrimination task (Supplementary Figure S10). Moreover, the absence of training and experience did not produce category effects during the preliminary JND measurements in the first group (Supplementary Figures S20a, S21b, and S22b). Finally, the comparison of JNDs between the two groups even indicated that categorical patterns were more pronounced in the post hoc than in the preliminary JND measurements (Supplementary Figure S19). The inverse would be expected if familiarity with the discrimination task counteracted categorical facilitation because the second and not the first group was experienced with the speeded discrimination experiment prior to the JND measurements. Consequently, training and familiarity alone are not sufficient to explain the absence of categorical facilitation effects in the first group.

Task sequence and task demands

The implementations of the discrimination task in the JND measurements and the speeded discrimination measurements differed in the presentation time (Figure 1), the stimulus sampling (Figure 3a), the blockwise presentation, and in the instructions and feedback. Our results indicate that JND measurements have a different effect on the succeeding speeded discrimination measurements (group 1) than speeded discrimination on succeeding JND measurements (group 2). The question arises how the sequence of the two kinds of discrimination measurements may affect the occurrence of categorical facilitation.

Our results suggest that the task for the JND measurements trained participants to perceive the color difference within the 500-ms presentation time in this task (cf. Supplementary Figure S1). This is supported by the observation that both groups become considerably faster during the JND measurements (Figure 7), and the first group answered at a response speed close to 500 ms at the end of the JND measurements and the beginning the speeded discrimination task (Figure 7a).

In contrast, the speeded discrimination task does not have that effect. Although the second group becomes faster during the speeded discrimination task (Supplementary Figure S8c), they do not reach the speed of the first group despite the explicit encouragement to respond as quickly as possible in this task (Figure 7b). Because this task shows colors until response, it does not require participants to see the color difference within a certain presentation time (Figure 1b). The instructions, the feedback, and the hall of fame that explicitly encouraged participants to respond as quickly as possible did not lead participants in the second group to answer as fast as the first group.

Hence, the presentation time constraint in the JND task trains observers to answer faster than what they judge themselves to be fastest when following the instructions in the speeded discrimination task. This indicates that observers learn through the presentation time constraint that 500 ms is sufficient to complete the task. This idea is consistent with the participant reports. Several participants reported that they responded by intuition rather than careful inspection in the JND measurements.

The observation that the second but not the first group yielded categorical facilitation may be explained by the distinction between slow and fast psychophysics (Schmidt et al., 2011). In both implementations of the discrimination task, the perceptual signal that allows for completing the task consisted of the hue difference between test and comparison. Such basic sensory information about color differences can be processed without visual awareness as shown for example by studies on blindsight (Stoerig & Cowey, 1989, 1991, 1992) or response priming (Schmidt, 2002). According to the idea of "fast psychophysics," visual signals are processed in a fast feed-forward, bottom-up direction through the visual system ("rapid chase"), similar to a "fast feed-forward sweep" (Lamme & Roelfsema, 2000). These processes may happen without visual awareness, for example, in subliminal priming (Schmidt, 2002).

Through the presentation time constraint, the JND measurements taught participants to rely on the sensory feed-forward color signal. As a result, they did not pay attention to the linguistic distinction between categories and completed the discrimination task based on their intuition rather than conscious inspection. In this way, the linguistic distinction between categories could not interfere with perceptual discrimination, at least not as much as in inexperienced observers.

In contrast, the inexperienced observers of the second group consciously inspected and evaluated the color differences in the speeded discrimination task. During the conscious analyses of the color differences, participants shifted their attention to aspects of the task they deemed to be important. In particular, due to their prior experience in everyday life, they might shift their attention to the categorical distinction. This shift of attention may happen automatically and without conscious decision or as a strategy that is part of the visual analysis. As a result, these observers, who did not learn to rely on the sensory signal prior to the speeded discrimination task, show categorical facilitation throughout all five sessions of the speeded discrimination task (Supplementary Figure S10).

This idea may also explain why the second group showed traces of category effects in the post hoc JND measurements (Supplementary Figure S19). The participants of the second group did not completely rely on the sensory signal as indicated by the fact that the response times in the post hoc measurements were not (yet) as low as those of the first group during the preliminary measurements (Figure 7). Instead, their tendency to direct their attention toward the categorical difference may have produced some traces of category effects in the post hoc JND measurements.

In sum, we propose that categorical facilitation occurs because observers pay attention to the linguistic distinction between categories. This idea explains the differences in categorical facilitation between the two groups. The untrained, inexperienced observers paid attention to the categorical distinction, and the trained observers followed their intuition about the sensory signal of the color difference. Future studies are necessary to test this idea and to establish the precise conditions that modulate effects of categorical facilitation.

Is there a lateralized category effect?

The idea of a lateralized category effect is supposed to reflect the impact of language on perception (Gilbert et al., 2006). According to this idea, category effects occur predominantly in the right visual field because it is processed by the left hemisphere, which, in turn, is processing language. In speeded discrimination, the idea suggests that communication between the hemispheres requires additional time, which would reduce category effects in the left visual field. Following this reasoning, observations of a right-lateralized category effect were taken to support the idea that language influences perception (in terms of discrimination performance) because the lateralized category effect reflects the lateralization of language processing. A wide range of studies have found evidence of such lateralization effects in behavioral (Gilbert et al., 2006; Drivonikou et al., 2007; Franklin, Drivonikou, Bevis, et al., 2008; Franklin, Drivonikou, Clifford, et al., 2008; Roberson et al., 2008; Roberson & Pak, 2009; Zhou et al., 2010; Paluy et al., 2011) and neurobiological measurements (Siok et al., 2009; Kwok et al., 2011; Mo, Xu, Kay, & Tan, 2011; Liu et al., 2010; A. Lu et al., 2014).

However, in a previous series of studies, we reimplemented the exact conditions that produced the patterns assumed to be lateralized category effects in the original studies (Witzel & Gegenfurtner, 2011). None of the 10 reimplementations with overall more than 200 observers yielded a lateralized category effect. Other studies used slightly different stimuli and procedures and also did not find a lateralization effect (Suegami, Aminihajibashi, & Laeng, 2014) or did not even find category effects (A. M. Brown et al., 2011; Lindsey et al., 2010). A recent study (Alvarez, Clifford, Holmes, & Franklin, 2012) showed that lateralization effects may occur when color differences and prior expectations about the side of the target location are insufficiently controlled, but these effects are unrelated to color categories.

The core problem in those previous studies was that the stimuli that were used to show the existence of a lateralized category effect poorly controlled for perceptual determinants of discriminability. All of the above studies only incompletely controlled color differences between stimulus pairs, mostly through Munsell steps or CIELUV distances. Moreover, most of these studies investigated the problematic green–blue boundary. The results of Witzel and Gegenfurtner (2011) suggested that the set of green–blue stimuli that has been used in those studies was prone to produce spurious category effects. For this reason, it was unclear whether the patterns that were previously interpreted as category effects were genuine category effects at all. Because those stimulus sets did not even allow for testing genuine category effects, they could not be used to prove the lateralization of category effects.

The present study solved the problem of perceptual differences by using equally discriminable colors. For the second group in our study, patterns of categorical facilitation still appeared for these stimuli in almost all categories, indicating genuine category effects. However, these genuine category effects in the second group were not lateralized either (Supplementary Figures S12, S13c and d, and S14c and d).

It might be objected that the colored disks in the present study were presented parafoveally. They were much narrower together than in the original studies that showed the lateralized category effect (Gilbert et al., 2006). Moreover, the fact that category effects in the second group appeared in both visual fields might be due to the high response times in this group (Roberson et al., 2008). Finally, failures to reveal significant lateralization effects may also be due to low statistical power, at least when testing across the nine participants of the first group. Hence, the few lateralization tendencies found in the first group might have become clear lateralized category effects with still another setup and more participants.

However, Witzel and Gegenfurtner (2011) showed that the exact settings of the original studies and large samples of participants did not reliably yield the expected lateralization effects. The present study did not find any lateralization effects for genuine category effects with equally discriminable colors. Together these findings raise serious doubts that there are any genuine lateralized category effects at all.

Conclusion

The present study investigated whether the linguistic distinction between color categories influences the performance in a speeded discrimination task. The particularity of this study was that color pairs were made equally discriminable through empirical discrimination thresholds. This approach allowed the resolution of ambiguities about the control of perceptual differences in previous studies.

Strong evidence for categorical facilitation was found for five out of six categories (pink, orange, yellow, green, and purple) with new, inexperienced observers (group 2). These effects were robust across individuals, response time distributions, time and experience, slight variations of category boundaries, and across visual fields. Because sensitivity to color differences was controlled through the use of equally discriminable colors, these findings exclude the possibility that these effects are methodological artifacts due to stimulus sampling. These findings show that the linguistic distinction between categories facilitates the discrimination of colors that coincide with the distinction between categories. Moreover, the absence of lateralized category effects in the present study casts further doubt on the existence of such effects.

In the blue category, discrimination performance was strongly modulated by factors that were not related to color categories. In particular, the cone-opponent L-M mechanism, i.e., a low-level early visual mechanism, seems to make the control of sensitivity in the vicinity of the green–blue boundary particularly difficult. With respect to previous studies that concentrated on the green–blue boundary, this finding highlights the difficulty of disentangling genuine category effects from variations in color sensitivity at that boundary.

Moreover, little categorical facilitation occurred in observers who participated in the extensive measurements of JNDs before completing the speeded discrimination task (group 1). These observers were highly trained and reached a level of performance that could not be achieved by the inexperienced observers of group 2. Additional analyses showed that the absence of robust categorical facilitation in these observers was not due to differences in color categorization, in response time distributions, to lateralization, or to the familiarity with the discrimination task. Instead, they suggest that categorical facilitation is strongly influenced by a combination of training and task demands.

These findings shed light on the mechanisms of categorical facilitation. They suggest that categorical facilitation occurs because naïve observers spontaneously pay attention to the linguistic distinction between categories. In contrast, highly trained observers learned how to distinguish the sensory signal about the color difference from noise without paying attention to the linguistic distinction between categories. This explains why inexperienced observers show categorical facilitation, but trained observers do not. Accordingly, attention to linguistic distinction between categories is at the origin of categorical facilitation. This idea opens a new path to understanding categorical perception beyond the realm of color.

Keywords: categorical perception, color categories, color cognition, color discrimination, color naming, language and perception, Sapir-Whorf hypothesis

Acknowledgments

We thank Walter Kirchner for technical assistance; Claudia Kubicek for assistance in data collection; and Sven Heinrich, Thomas Schmidt, Marie Rogers, and Carlijn van Alphen for helpful discussion. This research was supported by the German Science Foundation (DFG) Reinhart-Koselleck program Ge 879/9, Graduiertenkolleg GRK 885 "NeuroAct", and

SFB TRR 135, as well as by a Gießen University dissertation fellowship and a German Academic Exchange Service (DAAD) postdoctoral fellowship to C. W.

Commercial relationships: none.

Corresponding author: Christoph Witzel.
Email: cwitzel@daad-alumni.de.
Address: Laboratoire Psychologie de la Perception, Université Paris Descartes, Paris, France.

References

ActiveWire Inc. (2003). *ActiveWire* (Version 1.0.14). Palo Alto, CA: ActiveWire Inc. Retrieved from http://www.activewireinc.com/

Alvarez, J., Clifford, A., Holmes, A., & Franklin, A. (2012). Attention modulates hemispheric lateralisation of categorical colour search: An alternative account for 'Lateralised Whorf'. Paper presented at the Progress in Colour Studies 2012 (PICS12), Glasgow, UK.

Bachy, R., Dias, J., Alleysson, D., & Bonnardel, V. (2012). Hue discrimination, unique hues and naming. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 29*(2), A60–A68, doi:10.1364/JOSAA.29.000A60.

Bonnardel, V., van Leeuwen, C., & Flintham, J. (2007). Study of colour categorical perception with equal discriminability stimuli. *Perception, 36,* ECVP Abstract Supplement.

Bornstein, M. H., Kessen, W., & Weiskopf, S. (1976). The categories of hue in infancy. *Science, 191*(4223), 201–202.

Bornstein, M. H., & Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction times: Some implications for categorical perception and levels of information processing. *Psychological Research, 46*(3), 207–222.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436.

Brown, A. M., Lindsey, D. T., & Guckes, K. M. (2011). Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. *Journal of Vision, 11*(12):2, 1–21, doi: 10.1167/11.12.2. [PubMed] [Article]

Brown, R. W., & Lenneberg, E. H. (1954). A study in language and cognition. *Journal of Abnormal and Social Psychology, 49*(3), 454–462.

Cavonius, C. R., & Mollon, J. D. (1984). Reaction time as a measure of the discriminability of large colour differences. In C. P. Gibson (Ed.), *Colour coded vs monochrome electronic displays* (pp. 17.11–17.10). London: HMSO.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience, 13*(11), 1428–1432, doi:nn.2641 [pii] 10.1038/nn.2641.

Clifford, A., Franklin, A., Holmes, A., Drivonikou, V. G., Ozgen, E., & Davies, I. R. (2012). Neural correlates of acquired color category effects. *Brain and Cognition, 80*(1), 126–143, doi:10.1016/j.bandc.2012.04.011.

Collins, J. A., & Olson, I. R. (2014). Knowledge is power: How conceptual knowledge transforms visual cognition. *Psychonomic Bulletin & Review, 21*(4), 843–860, doi:10.3758/s13423-013-0564-3.

Cropper, S. J., Kvansakul, J. G., & Little, D. R. (2013). The categorisation of non-categorical colours: A novel paradigm in colour perception. *PLoS One, 8*(3), e59945, doi:10.1371/journal.pone.0059945.

Daoutis, C., Pilling, M., & Davies, I. (2006). Categorical effects in visual search for colour. *Visual Cognition, 14,* 217–240.

Davidoff, J., Davies, I. R. L., & Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature, 398*(6724), 203–204, doi:10.1038/18335.

Davidoff, J., Goldstein, J., Tharp, I., Wakui, E., & Fagot, J. (2012). Perceptual and categorical judgements of colour similarity. *Journal of Cognitive Psychology, 24*(7), 871–892, doi:10.1080/20445911.2012.706603.

De Valois, R. L., & De Valois, K. K. (1993). A multi-stage color model. *Vision Research, 33*(8), 1053–1065, doi: 0042-6989(93)90240-W [pii].

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in the lateral geniculate nucleus of macaque. *Journal of Physiology, 357,* 241–265.

Deutscher, G. (2011). *Through the language glass: Why the world looks different in other languages.* London: Random House UK.

Drivonikou, G. V., Kay, P., Regier, T., Ivry, R. B., Gilbert, A. L., Franklin, A., & Davies, I. R. L. (2007). Further evidence that Whorfian effects are stronger in the right visual field than the left. *Proceedings of the National Academy of Sciences, USA, 104*(3), 1097–1102, doi: 0610132104 [pii] 10.1073/pnas.0610132104.

Fahle, M. (2005). Perceptual learning: Specificity versus generalization. *Current Opinion in Neurobi-*

ology, 15(2), 154–160, doi:10.1016/j.conb.2005.03.010.

Fairchild, M. D. (1998). *Color appearance models.* Reading, MA: Addison-Wesley.

Fonteneau, E., & Davidoff, J. (2007). Neural correlates of colour categories. *Neuroreport, 18*(13), 1323–1327, doi:10.1097/WNR.0b013e3282c48c33 00001756-200708270-00005 [pii].

Forder, L., He, X., Witzel, C., & Franklin, A. (2014). Speakers of different colour lexicons differ only in post-perceptual processing of colour [Abstract]. *Perception, 43,* 33.

Franklin, A., Drivonikou, G. V., Bevis, L., Davies, I. R. L., Kay, P., & Regier, T. (2008). Categorical perception of color is lateralized to the right hemisphere in infants, but to the left hemisphere in adults. *Proceedings of the National Academy of Sciences, USA, 105*(9), 3221–3225, doi: 0712286105 [pii] 10.1073/pnas.0712286105.

Franklin, A., Drivonikou, G. V., Clifford, A., Kay, P., Regier, T., & Davies, I. R. L. (2008). Lateralization of categorical perception of color changes with color term acquisition. *Proceedings of the National Academy of Sciences, USA, 105*(47), 18221–18225, doi:0809952105 [pii] 10.1073/pnas.0809952105.

Gegenfurtner, K. R. (2003). Cortical mechanisms of colour vision. *Nature Reviews Neuroscience, 4,* 563–572.

Gegenfurtner, K. R., & Kiper, D. C. (2003). Color vision. *Annual Review of Neuroscience, 26*(1), 181–206.

Gellatly, A. (1995). Colourful Whorfian ideas: Linguistic and cultural influences on the perception and cognition of colour, and on the investigation of them. *Mind and Language, 10*(3), 199–225.

Gentner, D., & Goldin-Meadow, S. (2003). Whiter Whorf. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 3–14). Cambridge, MA: MIT Press.

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not in the left. *Proceedings of the National Academy of Sciences, USA, 103*(2), 489–494, doi: 0509868103 [pii] 10.1073/pnas.0509868103.

Goldstone, R. L., & Hendrickson, A. T. (2009). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(1), 69–78, doi:10.1002/wcs.26.

Gumperz, J., & Levinson, S. C. (1996). Introduction. In J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 1–20). Cambridge, UK: Cambridge University Press.

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (Chapter 1). New York: Cambridge University Press.

He, X., Witzel, C., Forder, L., Clifford, A., & Franklin, A. (2014). Color categories only affect post-perceptual processes when same- and different-category colors are equally discriminable. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 31*(4), A322–A331, doi:10.1364/JOSAA.31.00A322.

Heinrich, S. P., Kruger, K., & Bach, M. (2011). The dynamics of practice effects in an optotype acuity task. *Graefes Archives for Clinical and Experimental Ophthalmology, 249*(9), 1319–1326, doi:10.1007/s00417-011-1675-z.

Holmes, A., Franklin, A., Clifford, A., & Davies, I. R. L. (2009). Neurophysiological evidence for categorical perception of color. *Brain and Cognition, 69*(2), 426–434, doi:S0278-2626(08)00288-1 [pii] 10.1016/j.bandc.2008.09.003.

Huette, S., & McMurray, B. (2010). Continuous dynamics of color categorization. *Psychonomic Bulletin & Review, 17*(3), 348–354, doi:17/3/348 [pii] 10.3758/PBR.17.3.348.

Ishihara, S. (2004). *Ishihara's tests for colour deficiency*. Tokyo, Japan: Kanehara Trading Inc.

Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist, 86,* 65–79.

Kay, P., & Regier, T. (2006). Language, thought and color: Recent developments. *Trends in Cognitive Sciences, 10*(2), 51–54, doi:S1364-6613(05)00353-0 [pii] 10.1016/j.tics.2005.12.007.

Krauskopf, J., & Gegenfurtner, K. R. (1992). Color discrimination and adaptation. *Vision Research, 32*(11), 2165–2175.

Krauskopf, J., Williams, D. R., & Heeley, D. W. (1982). Cardinal directions of color space. *Vision Research, 22*(9), 1123–1131.

Kwok, V., Niu, Z., Kay, P., Zhou, K., Mo, L., Jin, Z., & Tan, L. H. (2011). Learning new color names produces rapid increase in gray matter in the intact adult human cortex. *Proceedings of the National Academy of Sciences, USA, 108*(16), 6686–6688, doi:1103217108 [pii] 10.1073/pnas.1103217108.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences, 23*(11), 571–579.

Laws, G., Davies, I., & Andrews, C. (1995). Linguistic structure and non-linguistic cognition: English and Russian blues compared. *Language and Cognitive Processes, 10*(1), 59–94.

Legge, G. E., & Foley, J. M. (1980). Contrast masking in human vision. *Journal of the Optical Society of America A, 70*(12), 1458–1471.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49*(2), 467–477.

Lindsey, D. T., Brown, A. M., Reijnen, E., Rich, A. N., Kuzmova, Y. I., & Wolfe, J. M. (2010). Color channels, not color appearance or color categories, guide visual search for desaturated color targets. *Psychological Science, 21*(9), 1208–1214, doi: 0956797610379861 [pii] 10.1177/0956797610379861.

Liu, Q., Li, H., Campos, J. L., Teeter, C., Tao, W., Zhang, Q., & Sun, H. J. (2010). Language suppression effects on the categorical perception of colour as evidenced through ERPs. *Biological Psychology, 85*(1), 45–52, doi:S0301-0511(10)00120-1 [pii] 10.1016/j.biopsycho.2010.05.001.

Lu, A., Yang, L., Yu, Y., Zhang, M., Shao, Y., & Zhang, H. (2014). Event-related potentials reveal linguistic suppression effect but not enhancement effect on categorical perception of color. *Scandinavian Journal of Psychology, 55*(4), 287–295, doi: 10.1111/sjop.12122.

Lu, Z. L., Hua, T., Huang, C. B., Zhou, Y., & Dosher, B. A. (2011). Visual perceptual learning. *Neurobiology of Learning and Memory, 95*(2), 145–151, doi: 10.1016/j.nlm.2010.09.010.

Lucy, J. A., & Shweder, R. A. (1979). Whorf and his critics: Linguistic and nonlinguistic influences on color memory. *American Anthropologist, 81*(3), 581–615.

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology, 3,* 54, doi:10.3389/fpsyg.2012.00054.

The MathWorks Inc. (2007). *Matlab - The language of technical computing* (Version R2007a). Natick, MA: The MathWorks Inc.

Mitterer, H., & de Ruiter, J. P. (2008). Recalibrating color categories using world knowledge. *Psychological Science, 19*(7), 629–634, doi:PSCI2133 [pii] 10.1111/j.1467-9280.2008.02133.x.

Mitterer, H., Horschig, J. M., Musseler, J., & Majid, A. (2009). The influence of memory on perception: It's not what things look like, it's what you call them. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35*(6), 1557–1562, doi:2009-19590-015 [pii] 10.1037/a0017019.

Mo, L., Xu, G., Kay, P., & Tan, L.-H. (2011). Electrophysiological evidence for the left-lateralized effect of language on preattentive categorical perception of color. *Proceedings of the National Academy of Sciences, USA, 108*(34), 14026–14030.

Mollon, J. D., & Cavonius, C. R. (1986). The discriminability of colours on c.r.t. displays. *Journal of the Institution of Electronic and Radio Engineers, 56*(3), 107–110.

Nagy, A. L., & Sanchez, R. R. (1990). Critical color differences determined with a visual search task. *Journal of the Optical Society of America A, 7*(7), 1209–1217.

Neitz, J., & Neitz, M. (2011). The genetics of normal and defective color vision. *Vision Research, 51*(7), 633–651, doi:10.1016/j.visres.2010.12.002.

Olkkonen, M., Witzel, C., Hansen, T., & Gegenfurtner, K. R. (2010). Categorical color constancy for real surfaces. *Journal of Vision, 10*(9):16, 1–22, doi:10.1167/10.9.16. [PubMed] [Article]

Özgen, E., & Davies, I. R. L. (2002). Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. *Journal of Experimental Psychology: General, 131*(4), 477–493.

Paluy, Y., Gilbert, A. L., Baldo, J. V., Dronkers, N. F., & Ivry, R. B. (2011). Aphasic patients exhibit a reversal of hemispheric asymmetries in categorical color discrimination. *Brain & Language, 116*(3), 151–156, doi:S0093-934X(10)00195-1 [pii] 10.1016/j.bandl.2010.11.005.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442.

Pilling, M., Wiggett, A., Özgen, E., & Davies, I. R. L. (2003). Is color "categorical perception" really perceptual? *Memory & Cognition, 31*(4), 538–551.

Raskin, L. A., Maital, S., & Bornstein, M. H. (1983). Perceptual categorization of color: A life-span study. *Psychological Research, 45*(2), 135–145.

Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition, 28*(6), 977–986.

Roberson, D., Davidoff, J., Davies, I. R. L., & Shapiro, L. R. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology,* *50*(4), 378–411, doi:S0010-0285(04)00076-3 [pii] 10.1016/j.cogpsych.2004.10.001.

Roberson, D., Davies, I. R. L., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General, 129*(3), 369–398.

Roberson, D., Hanley, J. R., & Pak, H. (2009). Thresholds for color discrimination in English and Korean speakers. *Cognition, 112*(3), 482–487, doi: S0010-0277(09)00139-5 [pii] 10.1016/j.cognition.2009.06.008.

Roberson, D., Pak, H., & Hanley, J. R. (2008). Categorical perception of colour in the left and right visual field is verbally mediated: Evidence from Korean. *Cognition, 107*(2), 752–762, doi: S0010-0277(07)00235-1 [pii] 10.1016/j.cognition.2007.09.001.

Roberson, D., & Pak, H. S. (2009). Categorical perception of color is restricted to the right visual field in Korean speakers who maintain central fixation. *Journal of Cognitive Science, 10*(1), 41–51.

Rosch Heider, E., & Olivier, D. C. (1972). The structure of the color space in naming and memory for two languages. *Cognitive Psychology, 3*(2), 337–354.

Rosenholtz, R., Nagy, A. L., & Bell, N. R. (2004). The effect of background color on asymmetries in color search. *Journal of Vision, 4*(3):9, 224–240, doi:10.1167/4.3.9. [PubMed] [Article]

Schmidt, T. (2002). The finger in flight: Real-time motor control by visually masked color stimuli. *Psychological Science, 13*(2), 112–118.

Schmidt, T., Haberkamp, A., Veltkamp, G. M., Weber, A., Seydell-Greenwald, A., & Schmidt, F. (2011). Visual processing in rapid-chase systems: Image processing, attention, and awareness. *Frontiers in Psychology, 2,* 169, doi:10.3389/fpsyg.2011.00169.

Siok, W. T., Kay, P., Wang, W. S. Y., Chan, A. H. D., Chen, L., Luke, K.-K., & Hai Tan, L. (2009). Language regions of brain are operative in color perception. *Proceedings of the National Academy of Sciences, USA, 106*(20), 8140–8145, doi:0903627106 [pii] 10.1073/pnas.0903627106.

Stoerig, P., & Cowey, A. (1989). Wavelength sensitivity in blindsight. *Nature, 342*(6252), 916–918, doi:10.1038/342916a0.

Stoerig, P., & Cowey, A. (1991). Increment-threshold spectral sensitivity in blindsight. Evidence for colour opponency. *Brain: A Journal of Neurology, 114*(Pt. 3), 1487–1512.

Stoerig, P., & Cowey, A. (1992). Wavelength discrimination in blindsight. *Brain: A Journal of Neurology, 115*(Pt. 2), 425–444.

Suegami, T., Aminihajibashi, S., & Laeng, B. (2014). Another look at category effects on colour perception and their left hemispheric lateralisation: No evidence from a colour identification task. *Cognitive Processing, 15*(2), 217–226, doi:10.1007/s10339-013-0595-8.

Valberg, A. (2001). Unique hues: An old problem for a new generation. *Vision Research, 41*(13), 1645–1657, doi:S0042-6989(01)00041-4 [pii].

Wilson, H. R. (1980). A transducer function for threshold and suprathreshold human vision. *Biological Cybernetics, 38*(3), 171–178.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences, USA, 104*(19), 7780–7785, doi:0701644104 [pii] 10.1073/pnas.0701644104.

Witthoft, N., Winawer, J., Wu, L., Frank, M., Wade, A., & Boroditsky, L. (2003). Effects of language on color discrimability. Paper presented at the 25th Annual Meeting of the Cognitive Science Society, Mahwah, NJ.

Witzel, C. (2011). Unterschiede in der Farbwahrnehmung [Translation: Differences in color perception]. In A. Groh (Ed.), *Was ist Farbe? - Bunte Beiträge aus der Wissenschaft* [Translation: *What is color? - Diverse contributions from science*] (pp. 39–62). Berlin: Weidler.

Witzel, C., & Franklin, A. (2014). Do focal colors look particularly "colorful"? *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 31*(4), A365–A374, doi:10.1364/JOSAA.31.00A365.

Witzel, C., & Gegenfurtner, K. R. (2011). Is there a lateralized category effect for color? *Journal of Vision, 11*(12):16, 1–25, doi:10.1167/11.12.16. [PubMed] [Article]

Witzel, C., & Gegenfurtner, K. R. (2012a). Category effects for red and brown [Abstract]. *Perception, 41,* 11.

Witzel, C., & Gegenfurtner, K. R. (2012b). No categorical appearance of equally discriminable colours. Paper presented at the Progress in Colour Studies 2012 (PICS12), Glasgow, UK.

Witzel, C., & Gegenfurtner, K. R. (2013). Categorical sensitivity to color differences. *Journal of Vision, 13*(7):1, 1–33, doi:10.1167/13.7.1. [PubMed] [Article]

Witzel, C., & Gegenfurtner, K. R. (2014). Category effects on colour discrimination. In W. Anderson, C. P. Biggam, C. A. Hough, & C. J. Kay (Eds.), *Colour studies: A broad spectrum* (pp. 200–211). Amsterdam: John Benjamin Publishing Company.

Witzel, C., Hansen, T., & Gegenfurtner, K. R. (2008). *Wie sich Farben mit den Betrachtern und mit den Zeiten ändern* [Translation: How colors change across observers and time]. Paper presented at the Tagung experimentell arbeitender Psychologen (TeaP), Marburg.

Witzel, C., Hansen, T., & Gegenfurtner, K. R. (2009). Categorical reaction times for equally discriminable colours [Abstract]. *Perception, 38,* 14.

Wuerger, S. M., Maloney, L. T., & Krauskopf, J. (1995). Proximity judgments in color space: Tests of a Euclidean color geometry. *Vision Research, 35*(6), 827–835, doi:0042-6989(94)00170-Q [pii].

Yokoi, K., Nishimori, T., & Saida, S. (2008). Interference of verbal labels in color categorical perception. *Optical Review, 15*(6), 295–301.

Yokoi, K., & Uchikawa, K. (2005). Color category influences heterogeneous visual search for color. *Journal of the Optical Society of America A, 22*(11), 2309–2317.

Zhou, K., Mo, L., Kay, P., Kwok, V. P. Y., Ip, T. N. M., & Tan, L. H. (2010). Newly trained lexical categories produce lateralized categorical perception of color. *Proceedings of the National Academy of Sciences, USA, 107*(22), 9974–9978, doi: 1005669107 [pii] 10.1073/pnas.1005669107.