Is there a lateralized category effect for color?

Christoph Witzel

Department of Psychology, University of Giessen, Giessen, Germany

Karl R. Gegenfurtner

Department of Psychology, University of Giessen, Giessen, Germany

According to the lateralized category effect for color, the influence of color category borders on color perception in fast reaction time tasks is significantly stronger in the right visual field than in the left. This finding has directly related behavioral category effects to the hemispheric lateralization of language. Multiple succeeding articles have built on these findings. We ran ten different versions of the two original experiments with overall 230 naive observers. We carefully controlled the rendering of the stimulus colors and determined the genuine color categories with an appropriate naming method. Congruent with the classical pattern of a category effect, reaction times in the visual search task were lower when the two colors to be discriminated belonged to different color categories than when they belonged to the same category. However, these effects were not lateralized: They appeared to the same extent in both visual fields.

Introduction

The *Lateralized Category Effect* is about the influence of language on perception. In the realm of color, this idea leads back to the question about the relationship between color perception and color categories. On the one hand, we perceive colors continuously in terms of hue, lightness, and saturation. On the other hand, when communicating about colors, we apply color names, such as green, blue, purple, etc. These color names refer to more or less discrete color categories that collapse the three dimensions of hue, lightness, and saturation. Evidence for category effects establishes a link between both color perception and color categories. In the case of a category effect, the presence of a categorical border between two colors accelerates their discrimination. This means that differences between two colors are identified faster when the colors belong to two different categories as compared to a color pair where both colors lie in the same category (e.g., Bornstein & Korda, 1984; Daoutis, Pilling, & Davies, 2006; Holmes, Franklin, Clifford, & Davies, 2009; Winawer et al., 2007; Witthoft et al., 2003). The reasoning behind the category effect is that the category border enhances the subjective appearance of color difference and, in this way, accelerates the deliberate identification of difference. This implies that category effects should also occur when the color differences to be detected are actually equivalent in terms of perceptual discriminability and only differ by the presence or absence of a category border (Witzel, Hansen, & Gegenfurtner, 2009).

A seminal paper has shown that this category effect is lateralized (Gilbert, Regier, Kay, & Ivry, 2006; henceforth

"Gilbert et al." will always refer to this article if not further specified). In the main task of this study, participants were shown an annulus of 12 colored squares. One of the squares was of different color than the others. Participants were asked to indicate whether the odd color was located on the left or right side, by pressing one of two keys. The color pairs to be discriminated were either across- or within-category color pairs. For across pairs, one of the two colors belonged to the green category and the other one to blue. The two colors of the within pairs belonged either both to green or both to blue. According to the idea of a category effect, the across pair should be discriminated faster than the within pair. Consequently, the localization reaction times should be lower for the across pair. This study, however, could show that this effect only occurs when the color to be discriminated was shown on the right instead of the left side of the screen. So the category effect was lateralized in that it only appeared in the right but not in the left visual field. Due to the contralateral projection of the visual pathways, the right visual field is connected to the left hemisphere (Purves, Augustine, & Fitzpatrick, 2004, pp. 263–267). Moreover, for almost all right-handers (and most left-handers), language areas are also localized in the left hemisphere (Knecht et al., 2000; Toga & Thompson, 2003; Tzourio, Crivello, Mellet, Nkanga-Ngila, & Mazoyer, 1998). Finally, lateralization in visual half-field tasks is a good predictor of language lateralization (Hunter & Brysbaert, 2008). These facts indicate that the lateralization of the category effect is related to the laterality of language. Additionally, Gilbert et al. have shown that the right lateralization of the category effect disappeared when observers performed a verbal interference task, which is

supposed to occupy the language areas in the brain. However, the effect persisted with a non-verbal interference task. These findings further strengthened the idea that the lateralization of the category effect was due to the influence of language areas on perception.

The discovery of the lateralization of the category effect opens up new groundbreaking research paths in psychophysics, on the one hand, and neurobiology, on the other (Masharov & Fischer, 2006; Regier & Kay, 2009; Roberson & Hanley, 2007). At the behavioral level, the lateralized category effect resolves ambiguities in the behavioral measures. One problem with the usual category effect has been to guarantee the comparability of the perceptual distances between the stimuli in each pair. Since the lateralized category effect is an interaction between the category effect and the visual field, the fact that stimuli within each pair are not perfectly equidistant does not undermine the authenticity of this effect. Another problem has been the question on whether color categories are really linguistic phenomena or rather the result of the particularities of the perceptual color space (Regier, Kay, & Khetarpal, 2007). The lateralization of the category effect, however, connects the category effect directly to language.

The psychophysical potential of this effect has been demonstrated by numerous follow-up papers. Drivonikou et al. (this short form will refer to Drivonikou, Kay et al., 2007) have shown the lateralization of the category effect in the data of Daoutis et al. (2006). Moreover, they used a procedure of Franklin, Pilling, and Davies (2005) that differed slightly from the one of Gilbert et al. They still found results that were consistent with the lateralized category effect. Further studies compared different language groups and found that the lateralized category effect appeared differentially at language-specific borders (Drivonikou, Davies, Franklin, & Taylor, 2007, Roberson et al., 2008; Roberson & Pak, 2009). Zhou et al. (2010) observed that participants could be trained to develop a lateralized category effect for artificial categories in as short a time as 3 h (see also Drivonikou, Clifford, Franklin, Özgen, & Davies, 2011). According to Franklin et al., the lateralization of the category effect switches during language acquisition from the right to the left visual field (Franklin, Drivonikou, Bevis et al., 2008; Franklin, Drivonikou, Clifford et al., 2008). Gilbert, Regier, Kay, and Ivry (2008) have claimed a general validity of the lateralized category effect beyond the realm of color perception by showing that it also appears for the discrimination of outline shapes of cats and dogs. For the categorization of oriented bars, an opposite lateralization effect in terms of accuracy has been found (Franklin, Catherwood, Alvarez, & Axelsson, 2010). However, the lateralization was reversed for infants, i.e., for orientation categories the infants' lateralization corresponded to the one for color categories in adults. While these studies did find lateralization to the left or right hemisphere, others did not find any lateralization (Liu, Chen, Wang, Zhou, & Sun,

2008), and some did not even find a category effect for color (Brown et al., 2009; Lindsey & Brown, 2009; Pinto, Kay, & Webster, 2010).

From the neurobiological perspective, the lateralized category effect links the neuroanatomy of language to behavioral effects of linguistic processing. Multiple neuro-imaging studies were motivated by the original works of Gilbert et al. and Drivonikou et al. The results were not conclusive. Tan et al. (2008) applied *functional magnetic resonance imaging* (fMRI) to show that color comparison resulted in the activation of language-specific brain areas as soon as the colors to be compared can be named easily. Likewise, using fMRI, Ikeda and Osaka (2007) observed a left hemispheric lateralization of color categorization. Kwok et al. (2011) found that 2 h of training with novel color categories produced an increase in gray matter in the left visual cortex (V2/V3) and in the right cerebellum (see their Figure 3). In contrast, Haslam et al. (2007) have shown that color categorization is barely changed in the progression of severe semantic dementia, implying that color categorization does not depend on the functionality of the language areas. Moreover, Fonteneau and Davidoff (2007) could not find any lateralization of brain activity involved in implicit color categorization when using *event-related potentials* (ERPs). Holmes et al. (2009) found a category effect on ERPs, which, however, was not lateralized. In a follow-up study, Clifford, Holmes, Davies, and Franklin (2010) observed that *visual mismatch negativity* (vMMN) was higher for color changes across than within color categories. This category effect only appeared in the upper but not in the lower visual field. However, recently Mo, Xu, Kay, and Tan (2011) found that this effect was also lateralized to the right visual field, indicating a lateralized category effect on ERPs.

Finally, three studies used slightly modified versions of the paradigm of Gilbert et al. to reveal neural correlates of the lateralized category effect. In an fMRI study, Siok et al. (2009) identified a left-side neuroanatomic lateralization that corresponded to the lateralization of the category effect in the right visual field. Liu et al. (2010, 2009) revealed a lateralization of category-specific ERPs. However, Siok et al. and Liu et al. (2009) could not or not unambiguously replicate the findings of Gilbert et al. at the behavioral level. Paluy, Gilbert, Baldo, Dronkers, and Ivry (2011) showed that the lateralization effect was reversed for aphasic patients with a stroke in the left hemisphere. At the same time, the two control groups and a group of right-hemisphere patients produced the original lateralized category effect.

Meanwhile, the lateralized category effect has also been introduced to a broader audience as a scientific proof for the influence of language on perception (e.g., Deutscher, 2011, p. 226 ff.; Hanley & Roberson, 2008; Kay, Regier, Gilbert, & Ivry, 2009). Given the high impact of the discovery of the lateralized category effect, it is of utter

importance that this effect may be reproduced. To achieve this, we reimplemented ten versions of the original experiments. Six of them reimplemented the procedure of the first experiment of Gilbert et al. and another four the one of Drivonikou et al. In our experiments, we paid particular attention on the rendering of colors. We simulated eight green–blue and three blue–purple stimulus sets on the basis of the specifications given in papers that found a lateralized category effect (Table S1 in the Supplementary material gives an overview of the different versions). In order to check whether reaction time differences may be due to differences of the distances between the single colors in each stimulus pair, we also measured discrimination thresholds for all the eleven sets of stimuli.

Study 1: Paradigm of Gilbert et al.

We tried six different versions of the first experiment of Gilbert et al. This experiment consisted of one part with the aforementioned visual search task alone and another part in which participants were given a verbal interference task during each block of the main task.

In the original experiment, there were four stimuli, referred to as A, B, C, and D. Colors A and B were categorized as green and C and D as blue. Consequently, the pairing of A and B or of C and D in the visual search task may be considered as within-category pairs. In turn, the pairing of colors B and C is an across-category pair since B belongs to the green category and C to the blue category (Gilbert et al., pp. 489–490). These four stimuli had *Munsell* hues of 7.5G, 2.5BG, 7.5BG, and 2.5B, respectively. "The brightness and saturation were adjusted to make them equal, based on the independent judgments of four observers (Gilbert et al., p. 493)." The resulting specifications for brightness (actually: lightness) and saturation were not given in the original article. Apart from this, the authors only provide the RGB values they used to render their stimuli and background on the computer screen (Gilbert et al., p. 493). However, this definition of the colors is device-dependent since the primaries may vary significantly across different monitors.

As a result, the color specifications in the original article are incomplete and do not allow an exact reproduction of the stimuli. To obtain a set of stimulus colors that may allow us to successfully replicate the lateralized category effect, we reimplemented four versions of the original experiment including the verbal interference task. In a first implementation (1.a), we simulated Munsell chips with constant lightness and saturation. In the second version (1.b), we used the RGB values specified by Gilbert et al. In the third and fourth versions (1.c and 1.d), we employed the green–blue and blue–purple stimulus sets as specified by Drivonikou et al.

We did not control for eye movements in those four versions of the experiment. Only when participants look at the center of the screen, the left side of the screen is seen in the left and the right side in the right visual field. Indeed, Roberson and Pak (2009) found a lateralization of the category effect for ten participants who maintained eye fixation to the center of the screen. However, when including the four participants who did not fixate properly, the category effect appeared on both sides. In order to control for eye movements, we ran the procedure of Gilbert et al. without verbal interference task and measured eye movements during the visual search task (1.e). For this purpose, we used the color specifications in CIE $L*u*v*$ reported in the latest articles on the lateralized category effect for English categories (Siok et al., 2009; Zhou et al., 2010). Our procedure allowed for comparing trials in which participants fixated the center and trials in which they did not. We repeated this experiment with an international sample of non-German participants (1.f) to exclude the possibility that our problem in replicating the lateralized category effect is due to particularities of the German color categories.

Methods

We concentrate on the basic characteristics of the main experiment. Supplementary information about the method may be found in the Method for main experiments section of the Supplementary material.

Participants

Participants were paid €8 per hour. All participants were monolingual and right-handers with a ratio above +0.6 in the Edinburgh Handedness Inventory (EHI; Oldfield, 1971). Color deficiencies have been excluded by means of the Ishihara Test (Ishihara, 2004). As in the original study, only participants "who placed the blue–green boundary between stimuli B and C were included in analysis of the visual search data" (Gilbert et al., p. 493).

In our version with the simulated Munsell chips (1.a), ten women and four men with an average age of 24.4 years (standard deviation = ±3.5 years) participated. In the version with the original RGB values (1.b), the sample consisted of thirteen women and two men with an average age of 23.1 ± 3.1 years. In the version with the green–blue color pairs of Drivonikou et al. (1.c), the sample consisted of 17 women and 3 men with an average age of 22.0 years ± 2.6 years. In the version with the blue–purple colors, 16 women and 4 men with an average age of 22.2 ± 3.3 years participated. For the implementation with the controlled eye movements, the German sample consisted of 14 women and 8 men with an average age of 23.1 ± 2.4 years. All participants in the aforementioned samples were native German speakers. The non-German sample consisted of 6 women and 3 men with an average age of 23.7 ± 3.8. Four of them were Italian, 3 were Spanish, 2 were English, and 2 of them were French. For further details,

see the Participants section in the Supplementary material.

Apparatus

To display the stimuli in the first four implementations, we used an *Iiyama MA203DT* monitor driven by an *NVIDIA* graphics card with a color resolution of 8 bits per channel. In the two implementations with the controlled eye movements, we used an *Eizo Color Edge CG223W-BK* monitor driven by an *NVIDIA Quadro FX1800* graphics card with a color resolution of 10 bits per channel. Monitors were calibrated and gamma corrected. Experiments were written in *MatLab* (The MathWorks Inc., 2007) with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). For the analysis of statistical power, we used the software G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007). In the first four implementations, responses were recorded by an ActiveWire device and in the other two by a wing-shaped game pad. In order to measure eye movements in the last two versions of the experiment, we used an *EyeLink II* (find further details in the Apparatus section of the Supplementary material).

Stimuli

Details on the conversion and rendering of the colors may be found in the Stimuli section of the Supplementary material. In particular, the chromaticity coordinates of the stimuli and the background are listed there, as measured on the monitor (Table S4).

In our first experiment (1.a), we rendered four *Munsell chips* with the hues specified in the original article by means of their reflectance spectra. In contrast to the original, we did not rely on observer judgments for the determination of saturation and lightness. Instead, we used a *Munsell value* of 5 and a *chroma* of 6 for all four stimuli. Value and chroma specify the lightness (relative brightness) and relative saturation in the *Munsell System*. By keeping them constant, the Munsell distance between two neighboring hues is the same for all four stimuli (Fairchild, 1998, pp. 115–117; Munsell Color Services, 2007). Moreover, the colors we specified are the same as the two-step stimuli that Bornstein and Korda (1984, p. 209) used in their seminal study on the category effect. We retrieved the reflectance spectra of the respective Munsell chips (7.5G5/6, 2.5BG5/6, 7.5BG5/6, 2.5B5/6) of the matte collection from the *Spectral database* of the University of Joensuu Color Group (2007). We converted the spectra into calibrated RGB values. The background was approximately standard illuminant C at half of the maximum monitor luminance. In our second implementation (1.b), we used the RGB values of the colors as given in the original article (Gilbert et al., p. 493). This implies that the rendered colors are different from the original

ones inasmuch as the primaries of the monitors are different. In these first two implementations, stimuli were rendered as squares of 2.1° of visual angle (24 mm at 655 mm distance).

Drivonikou et al. were the first to provide objective specifications of the colors they used to show the lateralized category effect. Hence, we also ran the procedure of Gilbert et al. using the stimuli specified by Drivonikou et al. Drivonikou et al. performed one experiment with green and blue and another one with blue and purple colors. In their experiment with green and blue, they had a set of four stimuli around the category border that were 5 hue steps apart in the Munsell System. The pairings of these stimuli will be called "far pairs." A second set of four stimuli was only 2.5 hue steps apart and was combined to "near pairs." In the blue–purple set, there were only four stimuli that were 2.5 hue steps apart. We employed the far green–blue stimuli in our third (1.c) and the blue–purple stimuli in our fourth (1.d) version of Gilbert et al.'s experiment. As a result, the stimuli of the green–blue set were distanced by 5 Munsell hue steps as in the experiment of Gilbert et al. However, they differed from the latter in that they were shifted by 2.5 hue steps toward green, and they did not involve any differences in chroma and value. We rendered these stimuli based on the CIE $L^*u'v'$ values given by Drivonikou et al. (p. 1101). To use these stimuli in the procedure of Gilbert et al., we rendered them on a gray background that corresponded to standard illuminant C at half of the monitor luminance. As in Drivonikou et al. (p. 1101), stimuli were shaped as disks with a diameter of ≈3.5° visual angle (40 mm at a distance of 655 mm).

For the version with the controlled eye movements (1.e–f), we rendered the CIE $L^*u^*v^*$ values given in Siok et al. (2009, p. 5; see also Zhou et al., 2010, p. 9977). Since there were slight deviations between the calculated and measured CIE 1931 chromaticity coordinates, we readjusted the stimulus colors by hand so that they corresponded exactly to the calculated chromaticity values.

For all five stimulus sets, each of the four stimuli (A, B, C, D) was paired with each other as in the original study (Gilbert et al., p. 489). As a result, there were overall six pairings (AB, BC, CD, AC, BD, AD).

Procedure

First, participants had to complete a naming pretest as described in Gilbert et al. (p. 493). Here, stimulus colors were shown in random order as squares on the gray background, and participants had to indicate whether the square was rather one or the other color category (green vs. blue or blue vs. purple).

The visual search task followed. In each trial, a fixation dot appeared followed by the ring of stimuli until a response was given. The participants had to indicate

whether the target was left or right by pressing a left or right key, respectively. As in the original article, the ring of stimuli was presented until a response was given. Then, a brief blank screen of 250 ms followed. As in Experiment 1 of Gilbert et al., in the first four implementations, there was a condition without any interference task and one with a verbal interference task. Each condition began with a respective practice block as described in the original paper. Then, three blocks with the main task followed. In each block, each of the two stimuli in a pair was once a target at each location on the ring. In the condition with verbal interference, the participants had to memorize eight digits in the correct order while doing the discrimination task. Before each block of the discrimination task, participants were shown eight random digits one after the other on the screen. Then, participants had to complete one of three blocks with the main task. At the end of such a block, eight underlines appeared on the screen and participants were to enter the eight memorized digits by means of a number pad.

When measuring eye movements, a fixation window (or region of interest, ROI in Roberson & Pak, 2009) was defined as 1.7 deg (2 cm) around the fixation cross. If participants moved their eyes out of the fixation window, the trial was considered as "non-fixated." Trials in which participants gave a wrong answer, responded slower than 1000 ms, or did not keep fixation were repeated later in the experiment. The same was true for every first trial after a drift correction or a calibration of the eye tracker.

Some parameters such as the duration of the fixation point or the number of possible positions varied across the different versions of the experiment. Detailed descriptions of these differences as well as of the specifications of the eye movement control are given in the Procedures section of the Supplementary material.

Data analysis

As in the original study, "[t]rials in which the participant pressed the wrong key […] or in which the reaction time was >2 *SD* from the participant's mean were not included in the analysis of the visual search data (Gilbert et al., p. 489)." Here, we focus on the within- and across-category pairs (1-step pairs in Gilbert et al., p. 490) and lump the two kinds of within-category pairs (AB and CD) together to compare them to the across-category pair (BC). If the reaction times for the across pair are lower than those for both within pairs together, we will call this the "classical pattern of the category effect." In order to test the occurrence of a lateralized category effect, we conducted a 2 (left vs. right visual field) × 2 (within- vs. across-category pair) Repeated Measurements Analysis of Variance (RMAOV) with reaction times as the dependent variable (Gilbert et al., p. 490). Moreover, we applied paired *t*-tests to analyze the visual fields separately. Further details on the data analysis for the versions with

the controlled eye movements may be found in the Data analysis for versions with controlled eye movements section of the Supplementary material.

Results

Contrary to the lateralized category effect, we found the classical pattern of the category effect in both visual fields. Moreover, the pattern of results was basically the same in the conditions with verbal interference as well as in trials without central fixation. In Figure 1, results are represented as in the original study (Gilbert et al., p. 490). The main results of the statistical analyses are given in Table 1. Further details are provided in the Results of main experiments section of the Supplementary material.

Main condition

Figure 1a depicts the original results of Gilbert et al. and the first row in Table 1a provides the corresponding statistical results. They found that in the right visual field reaction times for across pairs were on average about 24 ms lower than for within pairs; at the same time, there was no difference in reaction times between across and within stimuli in the left visual field (Gilbert et al., p. 490).

In contrast to these original results, we observed lower average reaction times for the across pair not only in the left but also in the right visual field. As can be seen in Figures 1c, 1e, 1g, 1i, 1l, and 1n (i.e., rows 2 to 7 in the left column), this is the case for all our implementations. In the implementation with the simulated Munsell chips, the reaction time difference between across and between pairs was highest, namely, 187 ms in the left and 205 ms in the right visual field. In the version with the non-German participants, this difference was smallest, namely, 19 ms in the left and 11 ms in the right visual field. In all our implementations, these differences were highly significant above zero in both visual fields ($P < 0.01$). As an exception to this, for the non-German participants, this difference was only significant in the left ($P < 0.05$) but not in the right visual field ($P > 0.18$). However, this result is opposite to the lateralized category effect and is probably spurious. There were no interaction effects in any of the conditions (see the last group of columns in Table 1a; all P at least 0.14).

Details about the main effects in the RMAOV may be found in Table S6a of the Supplementary material. As expected from the *t*-tests reported above, all main effects of pair type were highly significant (all $P < 0.01$). The only exception was the implementation with the non-German participants where this main effect was just marginally significant ($P < 0.10$). There was no main effect of visual field for our reimplementations either. Again, the only exception was the version with the non-German participants
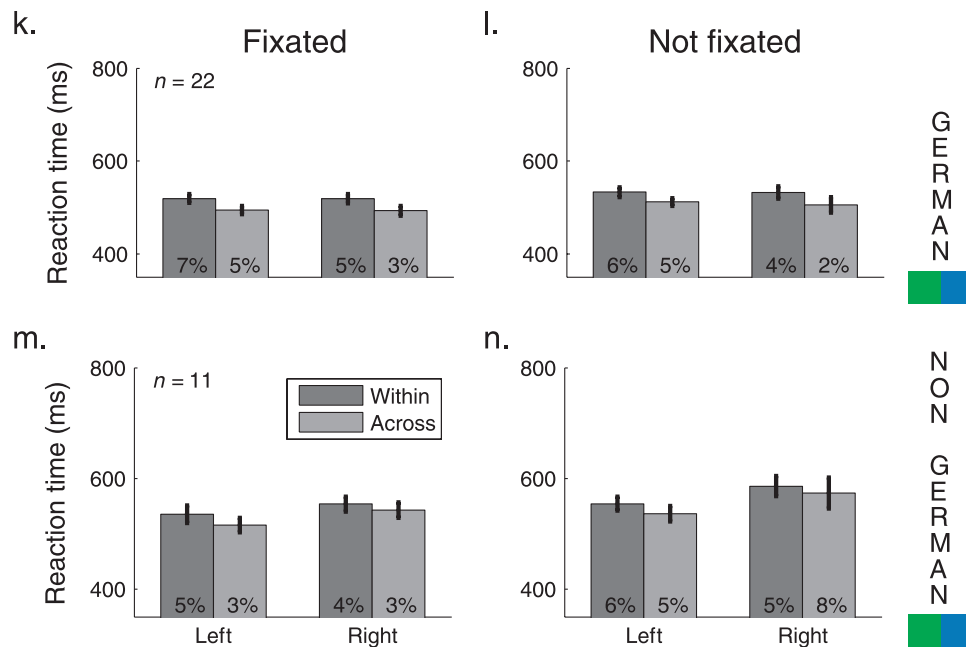
Witzel & Gegenfurtner
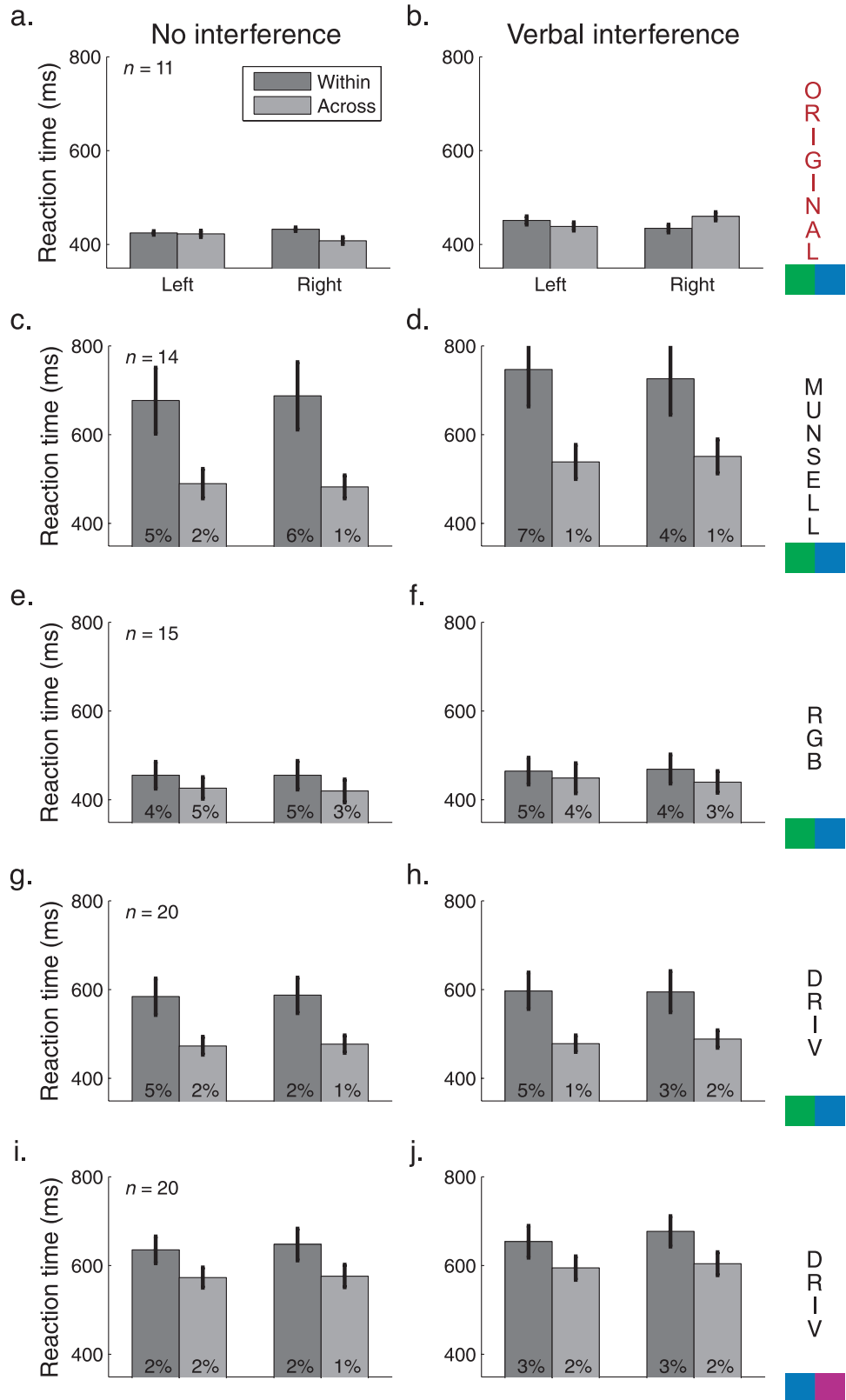
Figure 1. Average reaction times for implementations of Gilbert et al. Graphical representation as in Figure 1 of the original article (Gilbert et al., p. 490). Panels on the left side (a, c, e, g, i, k, and m) show reaction times for the main condition, in which the lateralized category effect was expected to occur. The panels on the right side represent the supplementary conditions. Note that the first five rows in the right column (b, d, f, h, and j) show results for conditions with verbal interference, in which the lateralized category effect should be disrupted. The last two rows depict the results for trials in which participants did not accurately fixate the center of the screen, which might also disrupt the lateralized category effect. Each row relates to one implementation as follows: (a, b) Original study of Gilbert et al. (p. 490, Figure 1); (c, d) our version 1.a with the simulated Munsell chips; (e, f) our version 1.b with the original RGB values; (g, h) our version 1.c with the green–blue stimuli of Drivonikou et al; (i, j) our implementation 1.d with the blue–purple stimuli of Drivonikou et al.; (k, l) results for the German participants in the implementation in which we controlled the fixation of the center; (m, n) those for the non-German participants. Dark bars depict the average reaction times for the within-category pairs, while light ones depict those for the across-category pairs. The left group of bars in each graphic concerns the left visual field (LVF), while the right one concerns the right visual field (RVF). As in the original article, error bars depict standard errors of mean (*SEM*). Numbers are error rates in percent. In all our implementations (c–n), the across-category pair yielded lower reaction times than the within-category pair independently of the visual field, verbal interference, and central fixation.

since they were faster on the right side. Apart from this exception, these results show that there was no systematic difference between response speed on the left and right sides despite the fact that all our participants were right-handers.

Condition with verbal interference

In four of our six versions of Gilbert et al.'s experiment, we included a condition with verbal interference. In none of these experiments, performance in the verbal interference task reached the 89% correct level reported in the original study of Gilbert et al. (p. 489). Performance varied between 67% and 79% correct, resulting in 37% and 14% completely correct blocks. For further details, see Table S5 in the Supplementary material.

In the original study, "the addition of the verbal-interference task reversed the visual field asymmetries observed when the visual search task was performed alone" (Gilbert et al., p. 490). This can be seen in Figure 1b.

Here, it is the left side, where the reaction times of across pairs were lower than those of within pairs. In the right visual field, the inverse was the case. In this way, the interaction between pair type and visual field in the RMAOV was reversed (cf. first row in Table 1b). Figures 1d, 1f, 1h, and 1j (i.e., rows 2 to 5 in the right column) show the results in our four versions. In contrast to the results of Gilbert et al., the profile of the reaction times in all our versions was similar to the condition without verbal interference. The across pair yielded lower reaction times than the within pair in both visual fields. Again, this difference was highest in our implementation with the simulated Munsell chips, namely, 208 ms in the left and 174 ms in the right visual field. It was smallest in the version with the original RGB values, namely, 16 ms and 29 ms, respectively. Given these small differences, it is no surprise that for this latter version the reaction time difference between across and within pairs was only marginally significant in the left visual field ($t(14) = 1.9$, $P = 0.08$). In all the other cases, these differences were

(a) Main condition

Study	df	Left W − A	t	P	Right W − A	t	P	Interaction LCE	F	P
Gilbert et al.	10	~0 ms	0.2	0.85	24 ms	2.8	*	~24 ms	16.1	**
Siok et al.	13	33 ms	3.9	**	45 ms	5.7	**	27 ms	3.6	°
Liu et al. (2009)[1]	11	40 ms	?	?	29 ms	?	?	−12 ms	2.0	0.19
Zhou et al.[2]	17	11 ms	−4.2	**	31 ms	−5.0	**	20 ms	8.3	**
Paluy et al.[3]	–	~25 ms	?	?	~45 ms	?	?	~20 ms	[7.1]	*
1.a Munsell	13	187 ms	4.1	**	205 ms	3.9	**	18 ms	2.4	0.14
1.b RGB	14	29 ms	3.2	**	35 ms	4.2	**	6 ms	1.6	0.22
1.c Green–blue	19	110 ms	4.7	**	110 ms	4.9	**	0 ms	0	0.99
1.d Blue–purple	19	63 ms	5.4	**	71 ms	5.1	**	8 ms	0.9	0.36
1.e German	21	25 ms	6.0	**	26 ms	4.6	**	1 ms	0.1	0.79
1.f Non-German	10	19 ms	2.3	*	11 ms	1.4	0.18	−9 ms	−1.2	0.27

(b) Supplementary conditions

Experiment	df	Left W − A	t	P	Right W − A	t	P	Interaction LCE	F	P
Gilbert et al.	10	13 ms	2.0	°	−26 ms	2.3	*	−39 ms	26.3	**
1.a Munsell	13	208 ms	4.0	**	174 ms	3.9	**	−33 ms	3.2	°
1.b RGB	14	16 ms	1.9	°	29 ms	3.0	**	13 ms	0.8	0.39
1.c Green–blue	19	119 ms	4.6	**	106 ms	3.5	**	−13 ms	1.5	0.23
1.d Blue–purple	19	59 ms	4.6	**	73 ms	6.1	**	16 ms	1.0	0.32
1.e German	21	21 ms	2.3	*	27 ms	2.2	*	6 ms	0.1	0.75
1.f Non-German	10	17 ms	1.4	0.18	12 ms	0.5	0.62	−5 ms	0.1	0.81

Table 1. Statistics for Study 1 on Gilbert et al. Statistical results for the condition (a) without and (b) with verbal interference. The rows correspond to different implementations of experiment to investigate the lateralized category effect as identified by the labels in the first column. For discussion, (a) includes the results of other available studies of the lateralized category effect apart from the one of Gilbert et al. Since they did not implement a condition with verbal interference, they do not appear in (b). The group of columns with the heading *Left* corresponds to the comparison of across and within pairs in the left visual field. *Right* corresponds to those in the right visual field, and *interaction* concerns the interaction between category and visual field, as indicative for the lateralized category effect. The degree of freedom within each factor is shown in the second column *df* and corresponds to $(n − 1)$, where n is the number of participants. $W − A$ refers to the difference in reaction time between within (W) and across pairs (A). A positive value refers to the classical pattern of the category effect, where responses to within pairs are slower than to across pairs. The columns t and P provide the results of a paired two-sided t-test across participants. The column LCE reports the size of the lateralized category effect. It is calculated as the difference between $W − A$ in the right and left visual fields. A positive value indicates that the classical pattern of the category effect is higher in the right visual field, as claimed by the proponents of the lateralized category effect. The columns F and P provide the F- and P-values of the two-way repeated measurement analysis of variance (RMAOV), with the factors 2 (category) × 2 (visual field). To provide a better graphical overview, the following symbols are used: ** = highly significant ($P < 0.01$), * = significant ($P < 0.05$), ° = marginally significant ($P < 0.1$), ? = information is missing; ~ = exact information is not available, but approximate information is available or may be inferred. *Notes*: [1]Liu et al. (2009) do not report results of separate t-tests for each visual field. [2]The specifications concern the pretraining measurement for the experimental group (Zhou et al., 2010, p. 9975). [3]The reaction time differences are taken from Paluy et al.'s Figure 2 for the 11 controls only; the results of the analysis of variance, though, are from a 3-way ANOVA that includes patients (with df = 16).

highly significantly above zero in both visual fields ($P <$ 0.01). Only for our first version, with the simulated Munsell chips, the interaction in the RMAOV was marginally significant ($F(1,13) = 3.2, P < 0.1$). There was a tendency toward a stronger pattern of the category effect in the left visual field (cf. rows 2 to 5 in Table 1b). For none of the other implementations, there was an interaction (minimum $P = 0.23$).

For details about the main effects in the RMAOV, see Table S6b of the Supplementary material. As expected from the t-tests reported above, all main effects of pair type were highly significant ($P < 0.01$). In the implementation with the blue–purple colors, there was a main effect for visual field, with faster responses in the left visual field ($F(1,19) = 5.6, P = 0.03$). In all other implementations, there was no main effect of visual field (minimum $P = 0.47$).

Trials without central fixation

Figures 1l and 1n show the reaction times for the German and the non-German sample of participants, respectively. The profile of reaction times is very similar to the ones we found in all the other experiments and conditions. The last two rows in Table 1b and in Table S6b of the Supplementary material report the results of the statistical analyses.

The number of non-fixated trials varied dramatically across participants. For the German sample, the total number of non-fixated trials per participant varied between a minimum of 19 and a maximum of 296, the average being 103.5. Despite this additional source of variability for the non-fixated trials, the statistics could reveal that the German participants responded faster to across-category pairs than to within-category pairs in both visual fields ($P < 0.05$). For the non-German participants, the average reaction times were also lower for the across than for the within pairs. However, as in the fixated trials, this difference was only significant in the left visual field and there was no significant main effect of category. Neither for the German nor for the non-German sample, there was an interaction between pair type and visual field (both $P > 0.75$). Finally, there was no main effect of visual field for the German sample of participants, but the non-German sample was faster in the right visual field ($P < 0.05$).

Overall performance

Figure 1 allows for an appreciation of the overall performance in terms of average reaction times and error rates by view. Details may be found in Table S5a in the Supplementary material.

In the original study of Gilbert et al. (p. 489), "[a]bout 8% of all trials were excluded by the criteria just mentioned, 75% of these because of erroneous responses"; this implies an error rate of about 6%. In all our reimplementations, error rates were lower than 6%. They were lowest in the implementation with the simulated Munsell chips (2%) and highest in the one with the controlled eye fixation (4%). Overall, in our versions, less than 8% of the trials had to be excluded. Exceptions were the implementations with controlled eye fixation, because eye fixation was an additional exclusion criterion that increased the amount of excluded trials. However, in all our versions, the proportion of error rates among the excluded trials was far less than the 75% reported by Gilbert et al.

Moreover, in our implementations, average reaction times were higher than in the original study. In the original study, average reaction times for the discrimination of 1-step distanced color pairs were 422 ms without verbal interference task and 444 ms with verbal interference task (Gilbert et al., p. 490). The version with the lowest reaction times among our experiments was the one with the original RGBs, where the average reaction times were 444 ms in the condition without verbal interference task and 459 ms in the condition with verbal interference task. The version with the simulated Munsell chips (613 and 666 ms) and the one with the blue–purple stimulus colors (618 and 642 ms) yielded highest average reaction times. As in the original study, in all our four experiments that included a condition with verbal interference, average reaction times were lower in conditions without verbal interference task than in those with the verbal interference task. In the two versions with eye movement control, average reaction times were higher in trials in which participants did not fixate.

Finally, Gilbert et al. (p. 489) reported that "[t]here was an approximately equal distribution of excluded trials between the two visual fields, and error rates were similar for within- and between-category trials." In Figure 1, the numbers at the bottom of the bars indicate the error rates in percent (note that the overall error rates mentioned above include 2- and 3-step pairs, which are not represented in the figure). For all our reimplementations, there was an approximately equal distribution of excluded trials between the two visual fields, too. However, there was a slight but consistent tendency of higher error rates when targets were presented on the left side. Moreover, there was a tendency of error rates to be lower for the across-category pairs than for the within-category pairs in all our implementations.

Discussion

In regard to the original hypotheses, all our experiments provided the same result: Instead of the lateralized category effect, we found the classical pattern of the category effect for both visual fields. Moreover, these reaction time patterns were robust to the verbal interference task in both visual fields and appeared whether participants maintained central fixation or not.

Since the original authors did not specify their stimulus colors correctly, we do not know the exact colors they used. The question arises whether our implementations differed from the original experiment in a way that prevented the effects found in the original study. Differences in the actual stimulus colors may change the location of the category boundaries, the relative distances of the color pairs, as well as the overall performance of the task. First, we discuss in how far the naming pattern of our participants fulfills the prerequisite that the category boundary lies between the center stimuli B and C of each stimulus set. Second, we analyze in how far the overall performance in our reimplementations was comparable with the one in the original as well as in other follow-up studies. Third, we will verify whether the performances across stimuli in our implementations reflect a genuine

category effect. Fourth, then, we discuss what might have prevented the lateralization of the category effect in the conditions in which it was expected. Finally, we will discuss why our implementations yielded almost the same results for the condition with verbal interference and for trials in which participants did not fixate the center.

Naming

The actual category border is a central prerequisite for the lateralized category effect. In the original study, 15% (2 of 13) of the participants were excluded because in the naming pretest their green–blue boundary did not coincide with the one assumed for the experiment. In some of our implementations, we observed a higher interindividual variability in the naming pretest than in the original study (cf. Pretest à la Gilbert et al. section in the Supplementary material). At the same time, several German and non-German participants reported that it was difficult to name the colors either green or blue since many of them were turquoise. In fact, Zimmer found some indications that, at least for German observers, there might be a turquoise region between blue and green (Zimmer, 1982; Zollinger, 1984).

With the pretest used in the original study of Gilbert et al., the presence of a turquoise boundary could not be noticed. This procedure forced participants to draw a border between green and blue by providing only these two response options. Moreover, the original procedure might induce the tendency to respond equally often with the different response options. If this was the case, the procedure would push the participants to set the boundary between stimuli B and C and artificially increase the consistency of the category border across participants.

We conducted several supplementary naming tests where people could choose freely among several color terms, including turquoise, to circumvent these problems (cf. Study 1 on Gilbert et al. section in the Supplementary material). In sum, these supplementary naming tests showed that the majority of participants tended to include turquoise when naming the green–blue colors, no matter whether they were German or not. The green–turquoise boundary lay between stimuli B and C and the turquoise–blue boundary between C and D of the green–blue stimulus sets. So, according to these results, there might be a supplementary boundary between C and D if we consider turquoise to be a relevant category for the evaluation of the main task. According to the supplementary tests for the blue–purple stimulus set, the blue–purple boundary was between A and B. This result contradicted the assumed blue–purple boundary between B and C. Though, this boundary has been confirmed with the original naming procedure. In sum, we may wonder whether the original naming pretest provided the genuine category borders and whether these category boundaries lay really between stimuli B and C in all our stimulus sets.

Overall performance

Our reimplementations yielded higher reaction times but also lower error rates than the original study. At first glance, the high reaction times together with the low error rates may indicate a speed–accuracy trade-off. In this case, our participants would try to avoid errors to the detriment of speed. This idea contrasts the explicit emphasis in the instructions to maximize speed instead of accuracy in our versions with the stimuli of Drivonikou et al. (1.c and 1.d).

However, not all variation in reaction times may be explained by a speed–accuracy trade-off. First, the average reaction times vary quite strongly across our different implementations, while error rates are equivalent. Moreover, neither the sizes of the reaction times nor of the error rates we measured are particular to our series of studies. In fact, the average reaction time of our second experiment in the condition without interference (444 ms) was lower than the one of Zhou et al. (465 ms), Liu et al. (467 ms), and Siok et al. (488 ms). At the same time, Zhou et al. also reported lower error rates (<6%) than the original, and the experiment of Liu et al. yielded even lower error rates (<2%) than any of our implementations (cf. Overall performance section in the Supplementary material).

Reaction times and stimulus similarity are directly related (e.g., Cavonius & Mollon, 1984; Mollon & Cavonius, 1986; Figure 3 in Nagy & Sanchez, 1990, p. 1212; Figures 4 and 5 in Rosenholtz, Nagy, & Bell, 2004, p. 228). The higher the similarity, i.e., the lower the perceptual distance, the higher the reaction times should be. We compared overall reaction times with the rank order of the Munsell distances, the ΔE_{Luv} distances, and the empirically measured discrimination thresholds across the aforementioned studies and our reimplementations (for details, see Perceptual distances and reaction times section in the Supplementary material). Surprisingly, only the rank order of the Munsell distances could clearly predict the reaction times ($r > -0.8$, $P < 0.01$). We conclude that the variability of reaction times across the different versions of the experiment is partly due to the difference in the overall perceptual distances of the respective stimulus sets.

Moreover, in almost all our reimplementations, the classical pattern of the category effect was more pronounced than in the study of Gilbert et al. However, except for our versions with the simulated Munsell chips (1.a) and with the stimuli of Drivonikou et al. (1.c and 1.d), our category effect patterns were in the range of those found by the follow-up studies of Siok et al., Liu et al., and Zhou et al. (cf. column 6 in Table S6). Given the difference in overall reaction times, the question arises of whether the size of the category effect pattern is related to the overall reaction times. Indeed, the pattern of the classical category effect is positively correlated to the overall size of reaction times ($r > 0.71$, $P < 0.05$; cf. Size of reaction times and

category effect section in the Supplementary material). If we assume reaction times to be an indicator for the difficulty of the task, then this correlation points toward a stronger classical pattern of the category effect for more difficult stimulus sets.

Performance across color pairs

If there was a genuine category effect, the across-category pair should not only be lower than the two within-category pairs together but should also be lower than each of the single within-category pairs. In the Study 1 on Gilbert et al. section in the Supplementary material, we thoroughly analyzed the differences of reaction times between the single color pairs. For all stimulus sets, we found that the reaction times for the assumed across pair (BC) were significantly lower than those for each within pair (AB and CD, respectively) in both visual fields separately and in both conditions (all $P < 0.05$). Only in the second implementation with the original RGB values and in the version with the non-German participants, this pattern was less pronounced. In the first case, the difference between BC and CD was not significant in the left visual field but in the right visual field (in both conditions; cf. Figures S1c and 1d). In the version with the non-German participants, it was the other way around. There was only one significant difference (namely, between AB and BC), and it appeared in the left visual field but not in the right visual field (cf. Figure S1k).

Moreover, for the green–blue stimuli of our versions 1.b and 1.c, the reaction time pattern of the main experiment was in line with the additional boundary between stimuli C and D we found in the supplementary naming test. Indeed, in both experiments, the assumed blue pair (CD) yielded significantly lower reaction times than the green pair (AB) in all conditions (P in the paired t-tests < 0.01, cf. Figures S1c–S1f). At the same time, the reaction times for this pair were significantly higher than for the green–blue pair (BC) as if the border between CD had a weaker effect on the reaction times. For the blue–purple stimuli of our implementation 1.d however, the data from the supplementary naming task were hard to reconcile with the reaction time data from the main experiment. The AB stimulus pair did not yield the lowest but the highest reaction times (cf. Figures S1g and S1h). The actual reaction time pattern for the blue–purple stimuli follows the originally assumed boundary between B and C.

However, the observed differences in reaction times across the stimulus pairs may also be due to differences in their discriminability. Differences in discriminability should affect performance in general. In this way, not only reaction times but also error rates should decrease with discriminability. In contrast, category effects do not necessarily depend on error rates in that they may also appear when accuracy is maximally high so that the error rates may not differ across stimuli because of the ceiling

effect. So, the concurrence of reaction times and error rates might point toward differences in discriminability. For this reason, we also inspected the distribution of error rates across the single stimulus pairs (cf. Study 1 on Gilbert et al. section in the Supplementary material). The coincidence of extremely high reaction times and error rates indicates that the blue stimulus pair (CD) in the first implementation with the simulated Munsell chips is much more difficult than the others. This was in line with some participants' reports after the experiment. In the implementations with the original RGB values and the two sets from Drivonikou et al., performance was least for the AB color pair. In the implementation with the controlled eye movements, low error rates coincided with the low reaction times of the across pair (BC).

For the implementation with the simulated Munsell chips (1.a), the discriminability measured as empirical JNDs may be an explanation for the lower reaction times of the across pair but, to our surprise, not for the particularly low performance of the blue stimulus pair (CD). For the implementation with the original RGBs and with the green–blue stimulus sets from Drivonikou et al., discriminability can well explain the lower performance of the green stimulus pair (AB). For both versions with the stimuli of Drivonikou et al. (1.c), it would predict the pattern of a category effect with lower reaction times for the across pair (BC). Finally, for the set of Siok et al., used in the implementations with the controlled eye movements, discriminability is not in line with the pattern of reaction times at all (see Perceptual distances and reaction times section in the Supplementary material).

In sum, for some of the implementations, discriminability may be a good explanation for the reaction time pattern. This might also provide an alternative explanation why the pattern of the category effect is correlated to the overall size of reaction times: Small variations in discriminability across stimulus pairs have a stronger impact on overall less discriminable stimulus sets. As a result, the observed differences in discriminability across stimulus pairs reinforce the pattern of a category effect more strongly for less discriminable stimulus sets, which also yield higher overall reaction times. However, at least the reaction times of the implementation with the original RGB values and of the implementations with the controlled eye movements may be better explained by the assumed green–blue boundary between B and C. Hence, these results are an indication of a genuine category effect.

Lateralization

Apart from the study of Gilbert et al., all four follow-up studies that used their procedure or a slightly modified version of it (Liu et al., 2009; Paluy et al., 2011; Siok et al., 2009; Zhou et al., 2010) found the classical pattern of the category effect in both visual fields (cf. Table 1). Paluy et al., Zhou et al., and Siok et al. found a significant or
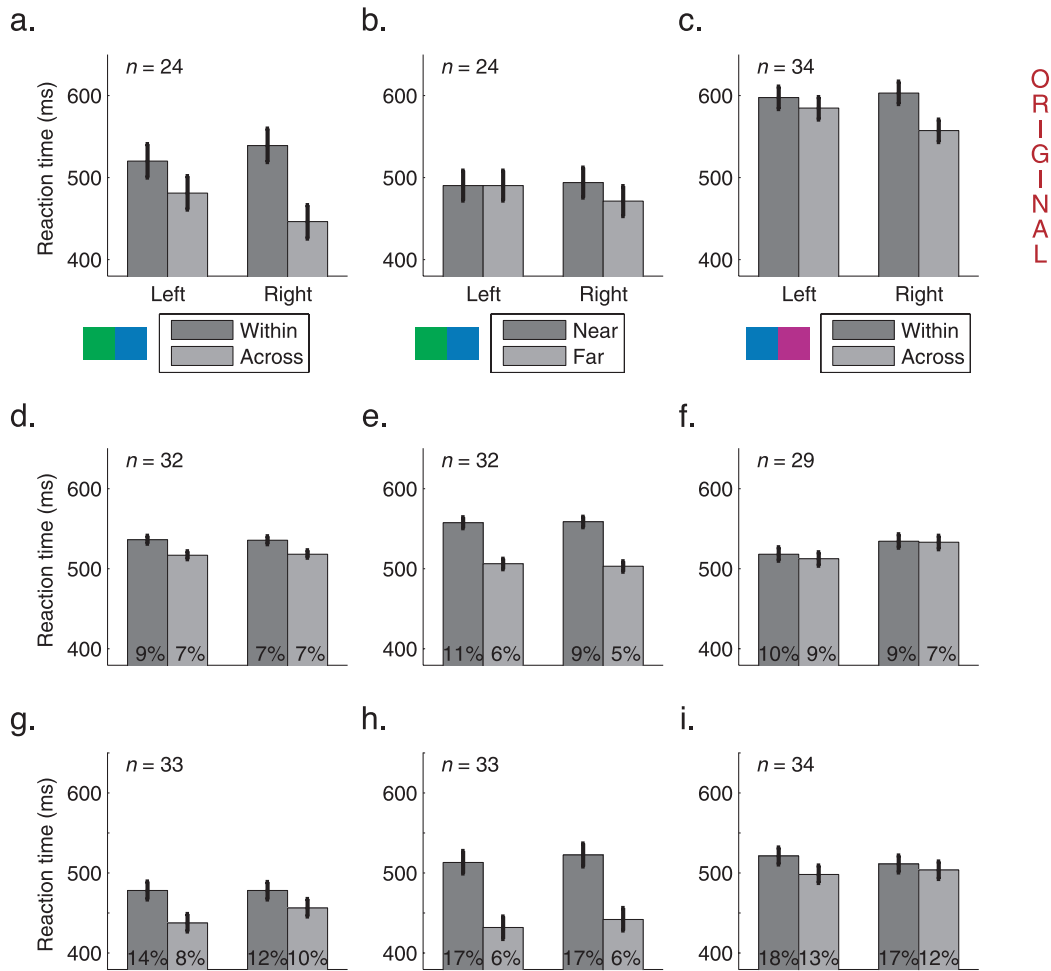
Figure 2. Average reaction times for implementations of Drivonikou et al. Graphical representation as in Figure 2 of the original article (Drivonikou et al., p. 1099). Bars are average medians, numbers are error rates in percent, and error bars correspond to confidence intervals. The calculations of the confidence intervals for our experiments were based on the pooled mean square error terms for the two-way comparisons as suggested by Masson and Loftus (2003, pp. 211–214). Panels on the left side allow for the evaluation of category effects in the experiments with green–blue stimuli. They do not differentiate between near- and far-distance stimuli. Panels on the right side allow for the evaluation of category effects in the experiments with blue–purple stimuli. Dark bars depict the average reaction times for the within-category pairs, while light ones depict those of the between-category pairs. Graphics in the middle differentiate between near- (dark bars) and far-distance stimuli (light bars) per visual field. Each row relates to one pair of experiments as follows: (a–c) the two original studies of Drivonikou et al.; (d–f) our simulations (2.a and 2.b); (g–i) our implementations with the original program (2.c and 2.d). The left group of bars in each graphic concerns the left visual field, while the right one concerns the right visual field. In all our reimplementations (d–i), the across-category pair yielded lower reaction times than the within-category pair independently of the visual field, and the far-distance pairs led to lower reaction times than the near-distance pairs.

marginally significant interaction that went into the direction of the lateralized category effect, but Liu et al. (2009) did not.

In our implementations, there were some small tendencies that may be interpreted as being in line with the hypotheses (cf. Table 1, column 3 from the right). In four of our six versions without verbal interference, the difference between across and within pairs was slightly higher in the right visual field. Furthermore, in two of these cases (1.a and 1.c), this relationship was reversed in the condition

with the verbal interference task in that this difference tended to be higher in the left visual field. In contrast, in the version with the original RGBs (1.b), this tendency was not reversed through the verbal interference task. Finally, in the version with the non-German participants (1.e), we even found a tendency in the main condition without interference that was opposite to the pattern of the lateralized category effect. However, none of all these tendencies yielded a statistically significant interaction between pair type and visual field.

The lack of statistical significance should not be due to the sample size since there were more participants in our implementations than in the original study (cf. Table S2 in the Supplementary material). Assuming an alpha and beta error of 0.05, all our reimplementations except the one with the non-German participants had enough statistical power in the main condition (i.e., no verbal interference, fixated) to detect small to medium effect sizes (>0.13 and <0.21) according to Cohen's (1988) criteria (cf. Power analyses section in the Supplementary material). So the question arises whether the performance in our implementations differed from the original experiment in a way that prevented the lateralization of the category effect. Two explanations are possible.

First, it has been argued that the absolute height of reaction times plays an important role for the lateralized category effect since long reaction times might enable cross-callosal transfer. According to this reasoning, the lateralized category effect only appears for very fast responses since otherwise there would be enough time for information to be communicated between the two hemispheres via the corpus callosum (Franklin et al., 2008; Regier & Kay, 2009; Roberson et al., 2008). However, we may observe that the average reaction times in all our versions were much lower than the ones Roberson et al. (2008, p. 759) considered to be reaction times of fast responders in their experiment (724 ms). If their fast reaction times were fast enough for a lateralization of the category effect, then the ones in our implementations should be fast enough anyway. Moreover, we did not find any correlation between the average reaction times and tendencies of lateralization, neither across experiments nor across individuals (cf. Size of reaction times and lateralization section in the Supplementary material).

Another possibility is that strong category effects supersede lateralization. The pattern of the category effect in our experiments may be too strong to allow for a lateralization. This is particularly true if this pattern is reinforced through differences in discriminability. In this case, lateralization may not be strong enough to completely prevent the pattern of the category effect in the left visual field. However, there was no correlation between the size of the category effect and its lateralization (cf. Size of category effect and lateralization section in the Supplementary material). This speaks against the idea that lateralization would have appeared for less pronounced patterns of category effects.

In sum, we could not find an overall explanation about why there was no lateralized category effect in our reimplementations. In some cases, the apparent pattern of the category effect might have been due to or enhanced by a higher discriminability of the across pairs. In turn, this pattern might have superseded any lateralization. This may have been the case in the first implementation with the simulated Munsell chips as well as in those with the stimuli of Drivonikou et al. However, in the implementa-

tion with the original RGBs and in those with the stimuli of Siok et al., there is no indication of how the characteristics of our implementations could have prevented the lateralization of the category effect.

Verbal interference and non-fixation

In our four reimplementations of the verbal interference task, the performance in digit memorization was much lower than in the original experiment (cf. Table S5a in the Supplementary material). The question arises whether these differences in performance were due to differences in implementation. In this regard, remember that we did include practice trials as in the original. However, some of our implementations of the verbal interference task differed slightly from the original (cf. Gilbert et al. with verbal interference task (1.a–d) section in the Supplementary material). Although, our version 1.c with the green–blue stimuli of Drivonikou et al. was designed according to supplementary information by the original authors, it still yielded a performance that was far from the one obtained by Gilbert et al. The low performance in our versions of the experiment might indicate that our participants did not engage sufficiently in these tasks. This may be the reason why these tasks did not yield the expected inversion effects, with a higher category effect in the left than in the right visual field.

Nevertheless, we are surprised that participants in the original experiment could remember all 8 digits in the correct order in 89% of all blocks. When passing the pilot versions of our implementations, we realized that this verbal interference task is very difficult. This has also been reported by almost all our participants. Since the working memory capacity is limited to about 5–9 items (Miller, 1956) or even less (Fukuda, Awh, & Vogel, 2010), this observation seems quite reasonable. Moreover, our results are completely in line with those of Liu et al. (2008). They used a one-back verbal interference task and found the classical pattern of the category effect in both visual fields, with and without verbal interference. However, even if we assume that the verbal interference task did not work in Liu et al.'s and all our versions, this cannot explain why we did not even find the lateralized category effect in the condition without interference.

Finally, it is no surprise that the trials in which participants look outside the fixation window in our last two experiments provided the same results as the trials with accurate fixation. After the training block, participants looked very rarely toward the differently colored stimulus. Instead, in many of the non-fixated trials, the participants' gaze drifted out of the tolerance region during the trial. In this way, the gaze was outside the fixation window at the end of these trials without being qualitatively different from the "fixated" trials. This implies that many of these trials were equivalent to the "fixated" trials, if we would have allowed a larger

tolerance. Hence, the results for these non-fixated trials are just another confirmation of the category effect independent of the visual fields.

Study 2: Paradigm of Drivonikou et al.

In our second series of studies, we reimplemented the experiments of Drivonikou et al. that employed the procedure of Franklin et al. (2005) to obtain the lateralized category effect. In our first two versions (2.a and 2.b), we simulated the experiments with the green–blue and blue–purple stimuli according to the information given in Drivonikou et al.'s article. To further adapt our implementations to the original ones, we reimplemented the two experiments a second time (2.c and 2.d). For this second attempt, the research group at the University of Surrey, UK, that conducted the original study of Drivonikou et al. provided us with supplementary information about the method and with the original program with which the original experiments were run (see Acknowledgment). Hence, in these versions, we used the original experimental program on our setup and adjusted all other parameters to match the ones of the original studies. As a result, there are several methodical differences in the exact procedure between our first (2.a and 2.b) and second (2.c and 2.d) implementations (for details, see Rationale behind study 2 on Drivonikou et al. section in the Supplementary material). In sum, while the first two versions of our reimplementations (2.a and 2.b) simulated the two original experiments according to the information given in the article, the second two versions (2.c and 2.d) were repetitions of the original experiments, only with new samples of participants.

Methods

As for the implementation of Gilbert et al. in the first part, further details on the method may be found in the Method for main experiments section in the Supplementary material.

Participants

Participants were paid €8 per hour. All participants were native German monolinguals speakers and right-handers with an EHI above +0.7 (Oldfield, 1971). Color deficiencies have been excluded by means of the Ishihara Test (Ishihara, 2004). In all our versions, the main task appeared to be very difficult. As a result, there were participants whose answers were close to chance (50% correct) or took more than 1 s to respond. This was even the case when using the original program and stimuli. In order to reduce outliers and to guarantee that reaction times are not too high, we excluded participants who gave less than 75% of correct answers or who needed on average more than 1000 ms to answer. The exclusion of participants did lower the average reaction times but did not change the profile of the results at all. In the simulation with the green–blue stimuli (2.a), the remaining sample consisted of 29 women and 3 men with an average age of 24.6 ± 4.2 years. In the simulation with the blue–purple stimuli (2.b), 26 women and 3 men participated (average age: 24.8 ± 4.4 years). Twenty-six women and 7 men (23.5 ± 3.7 years) took part in the version with the original program and the green–blue stimuli. Finally, the sample in the last implementation with blue–purple colors consisted of 30 women and 4 men (23.3 ± 3.7 years).

Apparatus

For the two simulations (2.a and 2.b), the apparatus was exactly the same as for the reimplementations of Gilbert et al. with the stimuli of Drivonikou et al. (1.c and 1.d, see first part). For the reimplementations with the original program, a *Sony GDM 20SE2T5* monitor was used together with an 8-bit *NVIDIA* graphics card. The program for these experiments was written in *Visual Basic 6.0*. As in the original study, the distance between observer and monitor was 50 cm, and answers were registered by a wing-shaped game pad.

Stimuli

Remember that Drivonikou et al.'s experiment with the green–blue stimuli contained four far-distance and four near-distance stimuli, here abbreviated as $A_fB_fC_fD_f$ and $A_nB_nC_nD_n$, respectively. Both are arranged so that the green–blue boundary, as verified in a naming pretest, was between the respective stimuli B and C. The near-distance across pair (B_nC_n) lay between the far-distance across pair. Hence, the two stimulus sets were nested as follows: A_f–A_n–B_f–B_n↔C_n–C_f–D_f–D_n (cf. Drivonikou et al., p. 1099, Figure 2a). The far-distance set was supposed to correspond—from green to blue—to the Munsell chips 10G7/8, 5BG7/8, 10BG7/8, and 5B7/8. So, they should be distanced by five Munsell steps. The near-distance set should correspond to 3.75BG7/8, 6.25BG7/8, 8.75BG7/8, and 1.25B7/8. Hence, this set implies pairwise distances of 2.5 Munsell steps. The blue–purple set was supposed to correspond—from blue to purple—to the Munsell specifications 6.25PB5/10, 8.75PB5/10, 1.25P5/10, and 3.75P5/10. So, as for the near-distance green–blue pair, the pairwise distances were 2.5 Munsell steps. However, Munsell specifications are only valid on a gray background N5. The white point of this background corresponds to standard CIE illuminant C and its lightness is equal to a Munsell value of 5 (Fairchild, 1998, p. 117; Newhall, 1940; Newhall, Nickerson, & Judd, 1943). Since the stimulus colors were not presented on a gray background,

it is not clear how much the actual color appearance of these colors deviated from the ones implied by the Munsell specifications. Certainly, the lightness of the colors was much higher than those of the Munsell specifications since in the original procedure participants mostly adapted to a black background (cf. Procedure section). In order to reproduce the exact color appearances of the stimuli, we took care to render the absolute colors used in the original study.

For the two versions that simulated the original experiment (2.a and 2.b), we computed the calibrated RGB values based on the CIE $L*u'v'$ specifications given in the original paper (Drivonikou et al., p. 1101). The target stimulus consisted of a disk with a diameter of 2.4° visual angle (27 mm on a distance of 655 mm). For the implementations with the original program (2.c and 2.d), stimuli were rendered in their absolute color, i.e., in the exact CIE 1931 chromaticity coordinates that correspond to the $Lu'v'$ values with the white point used by the original authors. These chromaticity coordinates have been double-checked with those of the original authors (Ian R. L. Davies, personal communication, February 25, 2009). The corresponding RGB values were first calculated and then adjusted manually so that they corresponded to the chromaticity coordinates as precisely as possible. The size of the target disk was 3.4° visual angle (30 mm with a distance of 500 mm).

As described in the original paper, each set of colors for green–blue far, green–blue near, and blue–purple were combined so that there was one pair within each of the respective categories and one across the respective categories. More precisely, the respective stimuli A, B, C, and D were combined to form the pairs AB, BC, and CD.

Procedure

Drivonikou et al.'s procedure differed from the one of Gilbert et al. as follows. In each trial of the main task, one of the two colors of a pair served as the target stimulus (i.e., the colored disk), while the other one was the background. This target could appear at one of 12 positions. The positions of the target were defined by 12 equally distant (30°) positions on a notional circle around the fixation point. The other color of the pair was shown as the background of the test display. Each trial began with the presentation of a white fixation marker on a black background for 1000 ms. Then, the test display was briefly flashed for 250 ms and people had to indicate on which side (left vs. right) the target was (cf. Drivonikou et al., p. 1101).

Data analysis

Following the original study, "[m]edian RTs for correct trials were calculated for each combination of category

(within/across), perceptual distance (near/far), and visual field (LVF/RVF) for each observer, and these data were subjected to a three-way repeated measures analysis of variance" (Drivonikou et al., p. 1100). For the green–blue stimuli, the factors were 2 (across- vs. within-*category pair*) × 2 (left vs. right *visual field*) × 2 (far vs. near *perceptual distances*). For the blue–purple stimuli, there were no far-distance stimulus pairs. Hence, the RMAOV had just two factors, namely, *category pair* and *visual field* (Drivonikou et al., p. 1100). We used again paired *t*-tests to analyze the effects of the single factors in detail.

Participants of our simulations (2.a and 2.b) completed more trials (864 and 432) than in the original study (96). Though, this is the same number of trials as in Gilbert et al., it might still be the case that in this particular procedure the lateralization of the category effect only occurs for the first few trials. For this reason, we will also report the results for the reduced data set of only the first 96 trials per participant for these two experiments.

Results

In both of our versions with green–blue stimuli (2.a and 2.b), there was a clear classical pattern of the category effect. Contrary to the lateralization hypothesis, this pattern appeared in both visual fields. There was only a non-significant tendency toward the interaction predicted by the original hypothesis when reducing the data set for the simulation (2.a) to the size of the data set in the original study. For the blue–purple stimuli, only the version with the original program (2.d) yielded a clear classical pattern of the category effect, and there were non-significant lateralization tendencies in the opposite direction of the lateralized category effect. For the main results, consider Figure 2 and Table 2; further details may be found in the Results of main experiments section of the Supplementary material.

Green–blue

Figure 2a and the first row of Table 2 recall the original results of Drivonikou et al.'s experiment with the green–blue stimulus set. Like us, they report a classical pattern of the category effect in both visual fields. However, in their experiment, the reaction time difference between across and within pairs is significantly stronger in the right visual field. The size of the lateralized category effect may be defined as the difference between the category effect in the right and left visual fields. In the original study, this size was about 60 ms for the ensemble of the green–blue stimuli. Moreover, they report that 22 of 24 (92%) observers showed the pattern of a lateralized category effect individually (Drivonikou et al., p. 1100).

For the corresponding results of our simulation and of the version with the original program, consider Figures 2d

Experiment	df	Left			Right			Interaction		
		$W - A$	t	P	$W - A$	t	P	LCE	F	P
Green–blue										
Drivonikou et al.	23	~30 ms[1]	>3.7	**	~90 ms[1]	>3.7	**	~60 ms	26.9	**
Liu et al. (2008)[2]	15	15 ms	?	?	20 ms	?	?	5 ms	–	–
Simulation	31	19 ms	9.2	**	17 ms	8.3	**	−2 ms	0.7	0.42
Simulation 96	31	13 ms	1.3	0.19	31 ms	2.3	*	18 ms	2.2	0.14
Original program	32	40 ms	5.9	**	22 ms	3.0	**	−19 ms	1.9	0.18
Blue–purple										
Drivonikou et al.	33	~15 ms	1.23	0.23	~45 ms	6.7	**	~30 ms	5.9	*
Simulation	28	6 ms	2.0	°	2 ms	1.6	0.69	−5 ms	1.2	0.29
Simulation 96	28	−13 ms	−0.4	0.72	−28 ms	−1.0	0.34	XX[3]	0.5	0.49
Original program	33	23 ms	3.5	**	8 ms	1.0	0.30	−15 ms	2.3	0.14

Table 2. Statistics for study 2 on Drivonikou et al. Rows and columns as well as symbols are the same as in Table 1. The only difference is that this table is divided in two parts, where the upper part (*green–blue*) reports the results for the implementations with a green–blue stimulus set and the lower part (*blue–purple*) reports those with a blue–purple stimulus set. The results for the green–blue stimulus sets lump together far- and near-distance pairs. Moreover, in the rows labeled *Simulation 96,* the table also reports the results for the reduced data set of our simulations with only the first 96 cases. *Notes*: [1]These values have been inferred from the information given in Drivonikou et al. (average category effect and lateralized category effect were both reported to be about 60 ms). [2]Liu et al. (2008) used the procedure of Drivonikou et al. together with the stimuli of Gilbert et al. They did only a three-way analysis of variance including a condition with verbal interference task as the third factor (p. 11). Hence, the statistics of the interactions are not directly comparable, and results of *t*-tests are not provided in the article. Moreover, this study only appears in the upper part of the table because it only used a green–blue stimulus set. This green–blue set does not contain near-distance stimulus pairs. [3]Since the reaction time pattern contradicted the category effect (see negative sign in columns 3 and 6), it does not make sense to report its lateralization.

and 2g (first column). For both versions, the paired *t*-tests revealed that in each visual field reaction times for the across-category pair were significantly lower than those for the across-category pairs (all *P* < 0.01, cf. Table 2). There was no significant interaction in the RMAOV (*P* = 0.42 and 0.18). Just 14 out of 32 (44%) and 12 of 33 participants (36%) yielded the pattern of a lateralized category effect in our simulation and the version with the original program, respectively. Only for the reduced data set of the simulation, these were more than 50%, namely, 20 out of 32 participants (63%). For the reduced data set, the difference between across- and within-category pairs was significant in the right visual field (*t*(31) = 2.3, *P* < 0.05) but not in the left visual field (*t*(31) = 1.3, *P* = 0.19). This is congruent with the original hypothesis. However, this tendency could not be confirmed by an interaction between category and visual field in the three-way RMAOV (*P* = 0.36, see Table 2 for details).

Figure 2b contrasts the reaction times for the two kinds of perceptual distances in each visual field as obtained by Drivonikou et al. In the original study, there was little difference between the reaction times for the far- and near-distance stimulus sets, and hence, there was no main effect of perceptual distance (Drivonikou et al., p. 1100). Our results are illustrated by Figures 2e and 2h. In our versions, the far-distance pair always resulted in lower reaction times—in the full and reduced data sets of the simulation as well as in the version with the original

program; in all cases, there was a main effect of perceptual distance in the RMAOV (all *P* < 0.01).

Finally, Drivonikou et al. (p. 1100) did not find a main effect of visual field. Instead, they obtained a yet unexplained interaction between perceptual distance and visual field, which may also be seen in Figure 2b. In contrast, we did not find any interaction but some main effects of visual field. For the reduced data set as well as for the version with the original program, the responses in the left visual field were faster. In the three-way RMAOV, these differences led to a highly significant and a marginally significant effect of visual field, respectively. There was neither another two-way nor a three-way interaction in any of our versions, neither with the full nor with the reduced data set (all *P* > 0.18).

Blue–purple

Figure 2c shows Drivonikou et al.'s results for the blue–purple stimulus set. They obtained the classical pattern of the category effect in the right visual field. On the left side, this pattern was 30 ms smaller and not statistically significant.

Figures 2f and 2i illustrate the reaction times we obtained with the simulation and the version with the original program, respectively. Contrary to the lateralized category effect, in both of these versions, the classical pattern of the category effect tended to be higher in the left visual field than in the right visual field. For the full data set of the simulation, the difference between the

reaction times for the across and within pairs was marginally significant in the left visual field ($t(28) = 2.0$, $P < 0.10$) but not in the right visual field ($t(28) = 0.4$, $P = 0.69$). Accordingly, there was not even a main effect of pair type in the two-way RMAOV ($F(1,28) = 1.8$, $P = 0.19$), and there was no interaction between pair type and visual field, either ($F(1,28) = 0.5$, $P = 0.59$). In the version with the original program, the classical pattern of the category effect was highly significant in the left visual field ($t(33) = 3.5$, $P < 0.01$) but again not significant in the right visual field ($t(33) = 1.0$, $P = 0.30$). For this version, there was also a main effect of category pair in the two-way RMAOV ($F(1,33) = 9.5$, $P < 0.01$). However, the lateralization tendency opposite to the lateralized category effect was not confirmed by an interaction between pair type and visual field ($F(1,28) = 1.2$, $P = 0.29$).

In the reduced data set, the profile of reaction times even contradicted the classical pattern of the category effect. The reaction times of the within pair were lower than for the across pair. However, these differences were neither confirmed by the single t-tests (both $P > 0.20$) nor by a main effect of pair type in the RMAOV ($F(1,28) = 0.45$, $P = 0.51$). There was no interaction between pair type and visual field ($F(33) = 2.3$, $P = 0.14$).

Finally, for all our versions, reaction times were lower in the left visual field than in the right visual field. Though, only for the full and reduced data sets of the simulation, this difference led to a significant main effect of visual field ($P < 0.01$ and $P < 0.05$); in the version with the original program, it did not ($P = 0.69$).

Overall performance

Consider Figure 2 and Table S5b in the Supplementary material to appreciate the overall performance in our versions of the experiment.

In the original study with the green–blue stimuli, the overall error rate was only 4% (Drivonikou et al., p. 1100) and the grand average of the median reaction times was about 490–500 ms (cf. Figure 2b or Figure 2d in Drivonikou et al.). In our two implementations with green–blue stimuli (2.a and 2.c), error rates were about twice or even thrice higher than in the original study. In regard to the reaction times, the simulation (2.a) yielded similar reaction times as the original study for the far-distance stimuli (504 ms) but clearly higher reaction times for the near-distance stimuli (558 ms; cf. Figures 2b and 2e). However, in this implementation, participants passed much more trials (864) than in the original study (96). Hence, they were more trained with the task and yielded increasingly lower reaction times in the course of the experiment. When considering the reduced data set (first 96 trials), the reaction times for the far- and near-distance green–blue stimuli (635 ms and 757 ms) were both much higher than those in the original study. In the implementation with the original program (2.c), reaction times were

lower for the far-distance (436 ms) and higher for the near-distance stimuli (517 ms) than in the original study.

For the blue–purple set, the average reaction time in the original study was about 590 ms (cf. Figure 2c or Figure 2e in Drivonikou et al.). Error rates were not reported for this experiment in the article. With the complete data set, the reaction times in our simulation with the blue–purple stimuli (2.b) were much lower (523 ms) than in the original study (cf. Figures 2c–2f). However, with the reduced data set, they were slightly higher than in the original study (609 ms). In our implementation with the original program, the average reaction times were much lower (511 ms) than in the original study (cf. Figures 2c–2i).

Discussion

Three of our four implementations of Drivonikou et al.'s experiments showed the classical pattern of the category effect: The across-category pairs yielded lower reaction times than the within-category pairs. In none of these experiments, there was a lateralized category effect. As for the reimplementations of Gilbert et al.'s experiment, we will discuss the naming pattern, the overall performance, the performance across stimuli, and the reasons for the absence of a lateralized category effect.

Naming

In the original study, the color naming pretest confirmed that the respective category boundary was between colors B and C of the respective stimulus set (Drivonikou et al., p. 1101). As in the first part, the answers to the naming pretests in our study varied across participants (cf. Pretest à la Drivonikou et al. section in the Supplementary material).

There are three possible reasons why the naming results of the original experiment might not apply to the results of our study. First, the original naming pretests were conducted with the colored squares on a gray background (Drivonikou et al., p. 1101), which was close to illuminant C (Ian R. L. Davies, personal communication, 2009). However, the actual background to which the participants adapted in the main task was black. Like color perception in general, color categorization does not depend on the absolute luminance but on the lightness of the colors (Shinoda, Uchikawa, & Ikeda, 1993; Uchikawa, Uchikawa, & Boynton, 1989). So, the question arises whether the original pretest is a valid test of the naming prerequisites for the main experiment. Second, given our naming results for the reimplementations of Gilbert et al. in the first part, the question arises in how far turquoise may play a role in categorization. Finally, all our participants in this second part and most of those in the first part were native German speakers, while those in the original studies were native

English speakers. Although the categories corresponding to basic color terms, such as green, blue, and purple, seem to be rather stable across languages of industrialized societies (e.g., Uchikawa & Boynton, 1987), there might be slight differences in the precise location of the category boundaries.

In order to verify the genuine category borders of the colors in the main task, we conducted several supplementary naming tests under the real conditions of the main experiment (cf. Study 2 on Drivonikou et al. section in the Supplementary material). Moreover, we compared the data of the categorization of Munsell chips by native German speakers (taken from Olkkonen, Witzel, Hansen, & Gegenfurtner, 2010) to the one for native English speakers as measured by Davidoff, Davies, and Roberson (1999; see the Munsell chip naming section in the Supplementary material for details). These analyses show that either the green–blue boundary is located slightly more toward blue than originally assumed, or there is a turquoise category and hence a supplementary boundary between turquoise and blue. However, the categorization in the region between green and blue also varies strongly across individuals. Since this variability does still increase when including turquoise, turquoise seems not to be equivalent to classical basic color terms. The deviation of our overall results from those of the original study may be due to the fact that our participants were native German speakers instead of native English speakers. However, given the naming pattern of the non-German participants in the first part, we may discard the idea that Germans have a particular naming pattern in the region between green and blue. For the blue–purple stimulus set, the naming under the valid conditions showed that the actual blue–purple boundaries lies rather between A and B and that there might be a supplementary purple–pink boundary between C and D. There is also no indication, however, that this pattern is a particularity of native German speakers. Rather, the observation of other results in the original study may just be due to the fact that they did not use the appropriate naming conditions.

Taken together, our supplementary investigations show that naming patterns deviate from the assumed ones when naming conditions are closer to those of the main experiment. These results undermine the validity of the original pretests for the main task. Moreover, it seems that the category boundaries are less clear-cut and stable than the original studies suggested.

Overall performance

The size of reaction times varied across the different versions, and error rates were higher in our versions than in the original study of Drivonikou et al. Remember that in our implementations only participants that yielded at least 75% of correct responses and an average reaction time below 1000 ms were included in the analyses.

Nevertheless, the participants' performance in our simulation with the green–blue stimuli (2.a) was much lower than the one in the original study. Possible explanations for the differences in performance across the implementations may be the different samples of participants or the different procedures. In the simulations (2.a and 2.b), the slightly different procedure might be the source of differences in performance (cf. Method for main experiments section in the Supplementary material). However, in our implementations with the original program, the procedure was exactly the same as in the original experiments. Nevertheless, it seems that error rates were much higher, while reaction times were slightly (green–blue) or clearly (blue–purple) lower than in the original study. Another possible explanation is a speed–accuracy trade-off, in which participants privileged speed over accuracy in our versions. However, even if this was the case, note that this speed–accuracy trade-off would be opposite to the one we found in our implementations (1.a–f) of Gilbert et al.'s study.

Furthermore, we already discussed in the first part that less similar colors should yield faster responses than more similar colors. Curiously, reaction times for far- and near-distance stimuli did not differ in the original study (Drivonikou et al., p. 1100). In all our implementations, the far-distance green–blue stimuli yielded lower reaction times than the near-distance stimuli. This was also true when comparing them to the blue–purple stimuli, which are, by definition, also near-distance stimuli. Moreover, the far-distance stimuli yielded lower error rates (cf. Figures 2e and 2h). These results are well in line with the assumption that near-distance (i.e., more similar) colors are more difficult to discriminate. Obviously, the high differences in performance between far- and near-distance stimulus pairs are reflected by their differences in discriminability. However, as with the reimplementations of Gilbert et al., only the rank order of the Munsell distances yielded some considerable correlations with average reaction times ($r \sim -0.75$, $P < 0.1$; cf. Perceptual distances section in the Supplementary material). For the coarse differences in perceptual distance across the experiments, Munsell distances can predict the size of reaction times, while ΔE_{Luv} and empirical JNDs cannot.

Finally, for the reimplementations of Gilbert et al. in the first part, we found a correlation between the size of reaction times and the strength of the classical pattern of the category effect. Here, there was no such correlation (cf. Size of reaction times and category effect section in the Supplementary material). This contradicts the idea that there is a simple and direct relationship between the pattern assumed to be a category effect and the overall size of reaction times. However, the absence of a correlation may also be due to differences in the pattern of discriminability across the stimulus sets or to other methodological variations across the different reimplementations.

Performance across color pairs

Remember that a genuine category effect would imply a dip of the reaction times for the across-category pair (BC). We observed that not all differences in reaction times follow this pattern of a genuine category effect. Moreover, not in all our implementations error rates were completely equally distributed across the different stimulus pairs. Again, the question arises whether some color pairs in a stimulus set were more difficult to discriminate and whether this may explain the results we observed.

In the Supplementary material (Study 2 on Drivonikou et al. section), we analyze the performance across stimulus pairs for each of the far- and near-distance sets separately. The pattern of reaction times was similar in the simulations (2.a and 2.b) and in the versions with the original program (2.c and 2.d). For all green–blue stimulus sets, the BC stimulus pair yielded lowest reaction times and error rates in both visual fields. At the same time, there was a tendency for the AB stimulus pair to yield highest reaction times and error rates in all green–blue sets. The only exception from this pattern was the near-distance green–blue set in the version with the original program (2.c). In this case, there was only a significant category pattern in the left visual field but no significant difference in the right visual field. Note that this lateralization tendency is opposite to the lateralized category effect and did not yield a significant interaction between category and visual field. This exception notwithstanding, these patterns of reaction times are in line with an impact of the assumed category boundary between B and C and a possible supplementary turquoise–blue boundary between C and D (cf. Category boundaries and reaction times section in the Supplementary material). However, for the far-distance but not for the near-distance stimuli, the difference in performance across stimuli could be predicted through the pattern of empirical JNDs. ΔE_{Luv} values failed to predict the pattern across stimuli (cf. Perceptual distances section in the Supplementary material).

Contrary to the prediction by a category effect, in both versions with a blue–purple stimulus set, there was no reaction time dip for the BC pair. Instead, the AB pair yielded the lowest performance in terms of reaction times and error rates, and the CD pair yielded the highest performance. This order of performance completely contradicts not only the pattern predicted by the category effect but also the one predicted through the measures of perceptual distance. Discrimination thresholds as well as ΔE_{Luv} distances would predict exactly the inverse profile of performance, highest for AB and lowest for CD. Taken together, discriminability and category boundaries fail to predict the low performance of the blue AB pair in these implementations as well as of the blue CD pair of the simulated Munsell chips in the first part (1.a). In the study of Brown, Lindsey, Rambeau, and Shamp (2009), blue shades were discriminated slower than green shades.

This observation might provide an explanation for the high reaction times of the blue pair. However, it is contradicted by the fact that in all other stimulus sets the blue pair yielded lower reaction times than the green pair.

Finally, as with the reimplementations of Gilbert et al., we could observe that the profile of empirical discriminability across stimulus pairs was different from the one of the ΔE_{Luv} distances. In particular, we observed a discrepancy between ΔE_{Luv} values and empirical JNDs in the region between green and blue. In particular, stimuli B and C of the green–blue stimulus sets seem to be more discriminable compared to the other stimulus pairs than predicted by the ΔE_{Luv} values. This observation is no surprise to us, since in the context of another study we found that there are particularly strong non-linearities in the green–blue region (Witzel, Hansen, & Gegenfurtner, 2008a). As a result, the empirical discrimination thresholds could predict the reaction time pattern even for stimulus sets for which Munsell distances were the same and ΔE_{Luv} distances were approximately equal. This shows that the fine-grained differences between the colors within each stimulus set may not be captured by the coarse measures of Munsell and ΔE_{Luv} distances.

In sum, there is no simple explanation for the patterns of reaction times across all our implementations. Some, but not all, of the patterns that were originally assumed to be category effects could be explained by the higher discriminability of the across pairs. Others seem to reflect the impact of the category boundaries and still others need further explanation, such as higher reaction times for blue. Maybe the overall explanation of the reaction time patterns has to be found in a combination of these factors. However, at this point, it is impossible to tease apart which factors exactly determined the profile of reaction times in each single stimulus set used here. Nevertheless, even if some of the reaction time patterns may have other origins, the ensemble of our results can be regarded as an indication of a genuine category effect. This is particularly true, when considering our findings in another series of studies. There we used equally discriminable color pairs and undistorted category borders and found evidence for genuine category effects for almost all chromatic basic color terms (Witzel et al., 2009).

Lateralization

None of our reimplementations of Drivonikou et al.'s procedure provided any lateralized category effect. There was only one tendency toward the lateralized category effect, namely, for the reduced data set in the simulation with the green–blue stimuli (2.a). In contrast, there were opposite lateralization tendencies in the two implementations with the blue–purple stimuli (2.b and 2.d) and for the near-distance green–blue stimuli of the version with the original program (2.c).

Again, it is unlikely that the lack of significant interaction effects was due to a lack of statistical power. In our implementations, the samples were much larger (green–blue) or equivalent (blue–purple) to the ones in the original study (cf. Table S2 in the Supplementary material). Assuming an alpha and beta error of 0.05, our versions of Drivonikou et al.'s experiments had enough power to detect small effect sizes (versions 2.a and 2.b) and small to medium effect sizes (versions 2.c and 2.d; find the details in the Power analyses section of the Supplementary material). We conclude that the statistical power is enough to detect an effect in at least one of our ten experiments.

Moreover, the lateralized category effect seems not to depend on the overall size of reaction times. In our last two implementations with the original program (2.c and 2.d), reaction times were lower than in the original study of Drivonikou et al., and in all our implementations, they were clearly lower than those reported by Roberson et al. (2008) for fast responders (724 ms). For this reason, we may exclude the idea that the lateralized category effect did not occur in our experiments because of the interhemispheric communication across the corpus callosum. Alternatively, if the lateralized category effect was bound to a high accuracy or to high reaction times, it should have appeared in one of our implementations of Gilbert et al., where reaction times and accuracy were comparatively high. However, it did not. Finally, as in the first part, there was no consistent relationship between the size of reaction times and potential lateralization effects, neither across individuals nor across experiments (cf. Drivonikou et al. section in the Supplementary material). So, we do not have any reason to believe that we would have obtained the pattern of a lateralized category effect with either lower or higher overall reaction times.

The present results also contradict the idea that the pronounced classical pattern of the category effect in our versions superseded its lateralization. In fact, in contrast to the results in the first part, here the classical patterns of the category effect were all lower in size than in the original study of Drivonikou et al. (<60 ms, cf. Table 2 and Table S7 in the Supplementary material). Moreover, across the different implementations, the classical pattern of the category effect is positively correlated with the lateralization pattern that has been interpreted as the lateralized category effect by the original authors (cf. Size of category effect and lateralization section in the Supplementary material). To reckon if strong category effects would supersede the lateralization effect, the correlation should be negative. However, it is not. Rather the correlation reflects the contrast between the patterns in the original studies and those of our implementations. As a result, we may definitely discard the idea of a superseding category effect pattern in our implementations.

Furthermore, neither for the implementations of Gilbert et al. nor for those of Drivonikou et al., there was a relationship between the lateralization tendencies of each individual participant and individual properties such as sex, age, handedness (size of EHI), or eye dominance (cf. Lateralization and participants section in the Supplementary material).

Finally, across the different studies of the lateralized category effect, there were several lateralization effects beyond the interaction between category effect and visual field. Drivonikou et al. (p. 1100) reported an interaction between visual field and distance, which they could not explain. In several of our reimplementations with different setups, procedures, and stimulus sets, we found reaction times to be lower in the left visual field than in the right visual field. In contrast, in the studies of Liu et al. (2009) and Siok et al. (2009) as well as in our implementation with the non-German sample, participants were faster in the right visual field. Given their heterogeneity, these lateralization effects seem unexplainable. Typically, one may expect that right-handers are faster with their right hand. Given the response mode of these experiments, participants should be faster on the right side. Some authors have argued that color detection is lateralized in the right hemisphere (e.g., Sasaki, Morimoto, Nishio, & Matsuura, 2007). This might facilitate responses in the left visual field. However, there is also evidence against the idea that color discrimination is lateralized (Danilova & Mollon, 2009). Finally, in each refreshment, the computer screen is built from left to right (and up to down). Maybe in procedures with brief stimulus presentations, such as the one of Drivonikou et al., this may systematically affect the stimulus timing and induce asymmetries across the screen. However, none of all these ideas may coherently explain the different lateralization effects found across the studies. For this reason, we even verified whether there may be slight color variations across the computer monitor. If such variations affected the relative discriminability of the stimulus pairs, they could be the source of asymmetries in performance across the screen. This could even produce spurious lateralized category effects. However, our measurements showed that spatial variations were barely stronger than the tiny variations of color rendering over time (cf. Asymmetries across the screen section in the Supplementary material).

Apart from the simulated Munsell chips and the original RGB values (versions 1.a and 1.b), our stimulus sets were very similar to or even exactly the same as the ones in the original studies. However, we did not find the pattern of the lateralized category effect. Instead, we observed that all kinds of lateralization effects may occur without any apparent connection to color categories. Moreover, we could show that the classical pattern of the category effect may well be due to the differential discriminability of these stimulus sets. Finally, the actual category borders of some of these stimuli are not even congruent with the pattern that was originally assumed to be a category effect. In these cases, one cannot expect any genuine lateralized category effect. Nevertheless, several studies, even beyond those of Gilbert et al. and Drivonikou et al.,

obtained the pattern of a lateralized category effect. Despite all our supplementary analyses, we are in lack of any explanation of why these studies found this lateralization pattern, while our systematic reimplementations as well as some other studies did not.

Conclusion

In ten different versions of the original experiments of Gilbert et al. and Drivonikou et al., we tried to replicate the lateralized category effect. Overall, we implemented two different procedures and employed eleven sets of stimulus colors. These stimulus sets differed in their discriminability (2.5, 5, and more than 5 Munsell steps) and used two different category boundaries (green–blue and blue–purple). In contrast to the original experiments, we carefully controlled our color rendering, we accounted for the observers' actual adaptation, and we determined the observers' genuine color categories for the actual lightness level. For all our sets of stimulus colors, our results exhibited the classical pattern of reaction times considered to be a category effect in the original studies. However, none of these effects were lateralized. They appeared in both visual fields, in conditions with and without a verbal interference task, and when participants maintain central fixation. A closer inspection of the results leads us to three conclusions.

First, we found that the naming patterns for the stimulus colors are less clear-cut and more complex than the original articles suggested. Although we observed a good general agreement between English and German basic color terms, we also found that the category membership of these colors may vary considerably between individual observers. These results are in line with previous findings that showed a considerable interindividual variability for color categorization in general (Hansen, Walter, & Gegenfurtner, 2007, Figure 2d; Olkkonen, Hansen, & Gegenfurtner, 2009, Figure 5; Webster et al., 2002) and for the categorization of colors close to the boundary in particular (Bornstein & Monroe, 1980, p. 218; Kay & McDaniel, 1978; Olkkonen et al., 2010, Figure 8; Witzel, Hansen, & Gegenfurtner, 2008b). In this context, we observed that the inclusion of turquoise as a supplementary category did rather increase than alleviate the indeterminacy of the category membership in the green–blue region.

Second, our findings at the same time support and undermine the existing evidence of a general category effect. For some of our stimulus sets, the differences in empirically measured discriminability provide an equally good or even better explanation than the category effect. This was even the case for some of the stimulus sets that were exactly the same as in the original studies. Hence, it is no wonder that studies that used sets of stimuli that differed from those of the original studies even failed to produce the pattern of a category effect (e.g., Brown et al., 2009; Lindsey & Brown, 2009; Pinto et al., 2010). Nevertheless, our implementations included also stimulus sets, where the reaction time pattern reflected the impact of the categories rather than the differences in discriminability. The ambiguity about the category effect is possible because research on category effects often confounds genuine category effects with discriminability. It has become a custom to prove the perceptual equivalence of color pairs by equal distances in the Munsell system or by equal Euclidean distances in CIELuv or CIELab space. That the perceptual uniformity of CIELuv, CIELab, and the Munsell system is only very coarse has been observed several times (Berns, 2000, pp. 107–130; Brainard, 2003, p. 206; Fairchild, 1998, p. 230; Wyszecki & Stiles, 1982, pp. 164–165). In particular, it has been known for long that distances in CIELuv are inappropriate to equate discriminability in reaction times (Cavonius & Mollon, 1984; Mollon & Cavonius, 1986). In our analyses, it has become clear that Munsell distances and CIELuv distances are too coarse to evaluate the fine-grained color differences that characterize the stimulus pairs in studies on the category effect. The quest for an adequate metric that guarantees the perceptual equidistance of color pairs remains one of the core challenges to research on categorical effects in color vision.

Third, according to the idea of the lateralized category effect, color categorization should be directly related to the left hemisphere. If there was such a direct relationship, reaction times for pure color naming should also be different across visual fields. This, however, is not the case (Bornstein & Monroe, 1980). Moreover, the neuropsychological evidence about the link between color categorization and cortical areas for language is contradictory (see Introduction section; for older studies, see discussion in Bornstein & Monroe, 1980, p. 217). Multiple studies looked for a lateralized category effect with diverse methodology but without success (Brown et al., 2009; Lindsey & Brown, 2009; Liu et al., 2008; Liu et al., 2009; Pinto et al., 2010; Siok et al., 2009). Our extensive series of experiments shows that the lateralized category effect could not be replicated despite systematic variations of the original experiments. Unfortunately, we cannot answer the question of which factors elicit or modulate the lateralization effects reported in the original and some of the follow-up studies. We could not replicate these results at all. In view of this outcome, the direct relationship between color categorization and hemispheric lateralization seems highly questionable to us.

Acknowledgments

Commercial relationships: none.
Corresponding author: Christoph Witzel.
Email: Christoph.Witzel@psychol.uni-giessen.de.
Address: Department of Psychology, University of Giessen, Otto-Behaghel-Straße 10F, 35394 Giessen, Germany.

References

Berns, R. S. (2000). *Principles of color technology* (3rd ed.). New York: John Wiley & Sons.

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,* 207–222.

Bornstein, M. H., & Monroe, M. D. (1980). Chromatic information processing: Rate depends on stimulus location in the category and psychological complexity. *Psychological Research, 42,* 213–225.

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

Brainard, D. H. (2003). Color appearance and color difference specification. In S. K. Shevell (Ed.), *The science of color* (pp. 191–216). Oxford, UK: Elsevier Science.

Brown, A. M., Lindsey, D. T., Rambeau, R. S., & Shamp, H. A. (2009). Visual search for colors as a test of the Sapir–Whorf hypothesis [Abstract]. *Journal of Vision, 9*(8):366, 366a, http://www.journalofvision.org/content/9/8/366, doi:10.1167/9.8.366.

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. 17.1–17.10). London: HMSO.

Clifford, A., Holmes, A., Davies, I. R. L., & Franklin, A. (2010). Color categories affect pre-attentive color perception. *Biological Psychology, 85,* 275–282.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Danilova, M. V., & Mollon, J. D. (2009). The symmetry of visual fields in chromatic discrimination. *Brain and Cognition, 69,* 39–46.

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,* 203–204.

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

Drivonikou, G. V., Clifford, A., Franklin, A., Özgen, E., & Davies, I. R. L. (2011). Category training affects colour discrimination but only in the right visual field. In C. P. Biggam, C. A. Hough, C. J. Kay, & D. R. Simmons (Eds.), *New directions in colour studies* (pp. 251–264). Amsterdam: John Benjamin Publishing Company.

Drivonikou, G. V., Davies, I. R. L., Franklin, A., & Taylor, C. (2007). Lateralisation of colour categorical perception: A cross-cultural study. *Perception, 36,* ECVP Abstract Supplement.

Drivonikou, G. V., Kay, P., Regier, T., Ivry, R. B., Gilbert, A. L., Franklin, A., et al. (2007). Further evidence that Whorfian effects are stronger in the right visual field than the left. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 1097–1102.

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

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191.

Fonteneau, E., & Davidoff, J. (2007). Neural correlates of colour categories. *Neuroreport, 18,* 1323–1327.

Franklin, A., Catherwood, D., Alvarez, J., & Axelsson, E. (2010). Hemispheric asymmetries in categorical perception of orientation in infants and adults. *Neuropsychologia, 48,* 2648–2657.

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 of the United States of America, 105,* 3221–3225.

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 of the United States of America, 105, 18221–18225.

Franklin, A., Pilling, M., & Davies, I. R. L. (2005). The nature of infant color categorization: Evidence from eye movements on a target detection task. *Journal of Experimental Child Psychology, 91,* 227–248.

Fukuda, K., Awh, E., & Vogel, E. K. (2010). Discrete capacity limits in visual working memory. *Current Opinion in Neurobiology, 20,* 177–182.

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 of the United States of America, 103,* 489–494.

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2008). Support for lateralization of the Whorf effect beyond the realm of color discrimination. *Brain Language, 105,* 91–98.

Hanley, J. R., & Roberson, D. (2008). Do infants see colors differently? *Scientific American Mind and Brain, May 14th 2008.*

Hansen, T., Walter, S., & Gegenfurtner, K. R. (2007). Effects of spatial and temporal context on color categories and color constancy. *Journal of Vision, 7*(4):2, 1–15, http://www.journalofvision.org/content/7/4/2, doi:10.1167/7.4.2. [PubMed] [Article]

Haslam, C., Wills, A. J., Haslam, S. A., Kay, J., Baron, R., & McNab, F. (2007). Does maintenance of colour categories rely on language? Evidence to the contrary from a case of semantic dementia. *Brain Language, 103,* 251–263.

Holmes, A., Franklin, A., Clifford, A., & Davies, I. R. L. (2009). Neurophysiological evidence for categorical perception of color. *Brain and Cognition, 69,* 426–434.

Hunter, Z. R., & Brysbaert, M. (2008). Visual half-field experiments are a good measure of cerebral language dominance if used properly: Evidence from fMRI. *Neuropsychologia, 46,* 316–325.

Ikeda, T., & Osaka, N. (2007). How are colors memorized in working memory? A functional magnetic resonance imaging study. *Neuroreport, 18,* 111–114.

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

Kay, P., & McDaniel, C. K. (1978). The linguistic significance of the meanings of basic color terms. *Language, 54,* 610–646.

Kay, P., Regier, T., Gilbert, A. L., & Ivry, R. B. (2009). Lateralized Whorf: Language influences perceptual decision in the right visual field. In J. W. Minett & W. S.-Y. Wang (Eds.), *Language, evolution, and the brain* (pp. 261–284). Hong Kong: The City University of Hong Kong Press.

Knecht, S., Dräger, B., Deppe, M., Bobe, L., Lohmann, H., Flöel, A., et al. (2000). Handedness and hemispheric language dominance in healthy humans. *Brain, 123,* 2512–2518.

Kwok, V., Niu, Z., Kay, P., Zhou, K., Mo, L., Jin, Z., et al. (2011). Learning new color names produces rapid increase in gray matter in the intact adult human cortex. *Proceedings of the National Academy of Sciences of the United States of America, 108,* 6686–6688.

Lindsey, D. T., & Brown, A. M. (2009). Color difference scaling at the blue–green color category boundary as a test of the Sapir–Whorf Hypothesis [Abstract]. *Journal of Vision, 9*(8):340, 340a, http://www.journalofvision.org/content/9/8/340, doi:10.1167/9.8.340.

Liu, Q., Chen, A.-T., Wang, Q., Zhou, L., & Sun, H.-J. (2008). An evidence for the effect of categorical perception on color perception. *Acta Psychologica Sinica, 40,* 8–13.

Liu, Q., Li, H., Campos, J. L., Teeter, C., Tao, W., Zhang, Q., et al. (2010). Language suppression effects on the categorical perception of colour as evidenced through ERPs. *Biological Psychology, 85,* 45–52.

Liu, Q., Li, H., Campos, J. L., Wang, Q., Zhang, Y., Qiu, J., et al. (2009). The N2pc component in ERP and the lateralization effect of language on color perception. *Neuroscience Letters, 454,* 58–61.

Masharov, M., & Fischer, M. H. (2006). Linguistic relativity: Does language help or hinder perception? *Current Biology, 16,* R289–R291.

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57,* 203–220.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

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 of the United States of America, 108,* 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,* 107–110.

Munsell Color Services. (2007). *The Munsell book of color—Glossy collection.* Grandville, MI: X-rite.

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,* 1209–1217.

Newhall, S. M. (1940). Preliminary report of the O.S.A. subcommittee on the spacing of the Munsell colors. *Journal of the Optical Society of America, 30,* 617–645.

Newhall, S. M., Nickerson, D., & Judd, D. B. (1943). Final report of the O.S.A. subcommittee on the spacing of the Munsell colors. *Journal of the Optical Society of America, 33,* 385–418.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9,* 97–113.

Olkkonen, M., Hansen, T., & Gegenfurtner, K. R. (2009). Categorical color constancy for simulated surfaces. *Journal of Vision, 9*(12):6, 1–18, http://www.journalofvision.org/content/9/12/6, doi:10.1167/9.12.6. [PubMed] [Article]

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

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,* 151–156.

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

Pinto, L., Kay, P., & Webster, M. A. (2010). Color categories and perceptual grouping [Abstract]. *Journal of Vision, 10*(7):409, 409a, http://www.journalofvision.org/content/10/7/409, doi:10.1167/10.7.409.

Purves, D., Augustine, G. J., & Fitzpatrick, D. (2004). *Neuroscience* (3rd ed.). Sunderland, MA: Palgrave Macmillan.

Regier, T., & Kay, P. (2009). Language, thought, and color: Whorf was half right. *Trends in Cognitive Sciences, 13,* 439–446.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 1436–1441.

Roberson, D., & Hanley, J. R. (2007). Color vision: Color categories vary with language after all. *Current Biology, 17,* R605–R607.

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*(Special Issue: Color in Thought and Language), 41–51.

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,* 752–762.

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, http://www.journalofvision.org/content/4/3/9, doi:10.1167/4.3.9. [PubMed] [Article]

Sasaki, H., Morimoto, A., Nishio, A., & Matsuura, S. (2007). Right hemisphere specialization for color detection. *Brain Cognition, 64,* 282–289.

Shinoda, H., Uchikawa, K., & Ikeda, M. (1993). Categorized color space on CRT in the aperture and the surface color mode. *Color Research & Application, 18,* 326–333.

Siok, W. T., Kay, P., Wang, W. S. Y., Chan, A. H. D., Chen, L., Luke, K.-K., et al. (2009). Language regions of brain are operative in color perception. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 8140–8145.

Tan, L. H., Chan, A. H. D., Kay, P., Khong, P.-L., Yip, L. K. C., & Luke, K.-K. (2008). Language affects patterns of brain activation associated with perceptual decision. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 4004–4009.

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

Toga, A. W., & Thompson, P. M. (2003). Mapping brain asymmetry. *Nature Reviews Neuroscience, 4,* 37–48.

Tzourio, N., Crivello, F., Mellet, E., Nkanga-Ngila, B., & Mazoyer, B. (1998). Functional anatomy of dominance for speech comprehension in left handers vs. right handers. *Neuroimage, 8,* 1–16.

Uchikawa, H., Uchikawa, K., & Boynton, R. M. (1989). Influence of achromatic surrounds on categorical perception of surface colors. *Vision Research, 29,* 881–890.

Uchikawa, K., & Boynton, R. M. (1987). Categorical color perception of Japanese observers: Comparison with that of Americans. *Vision Research, 27,* 1825–1833.

University of Joensuu Color Group. (2007). *Spectral database.* Available from http://spectral.joensuu.fi/.

Webster, M. A., Webster, S. M., Bharadwaj, S., Verma, R., Jaikumar, J., Madan, G., et al. (2002). Variations in normal color vision: III. Unique hues in Indian and United States observers. *Journal of the Optical Society of America A, 19,* 1951–1962.

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 of the United States of America, 104,* 7780–7785.

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

Witzel, C., Hansen, T., & Gegenfurtner, K. R. (2008a). Categorical discrimination of colour [Abstract]. *Journal of Vision, 8*(6):577, 577a, http://www.journalofvision.org/content/8/6/577, doi:10.1167/8.6.577.

Witzel, C., Hansen, T., & Gegenfurtner, K. R. (2008b). *Wie sich Farben mit den Betrachtern und mit den Zeiten ändern*. 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. *Perception, 38,* ECVP Abstract Supplement, 14.

Wyszecki, G., & Stiles, W. S. (1982). *Color science: Concepts and methods, quantitative data and formulae* (2nd ed.). New York: John Wiley & Sons.

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 of the United States of America, 107,* 9974–9978.

Zimmer, A. C. (1982). What really is turquoise? A note on the evolution of color terms. *Psychological Research, 44,* 213–230.

Zollinger, H. (1984). Why just turquoise? Remarks on the evolution of color terms. *Psychological Research, 46,* 403–409.