SUPPLEMENTARY MATERIAL

Categorical facilitation with equally discriminable colors

Christoph Witzel, & Karl R. Gegenfurtner

Table of Content:

METHOD

Response times at threshold (Fig. S1)

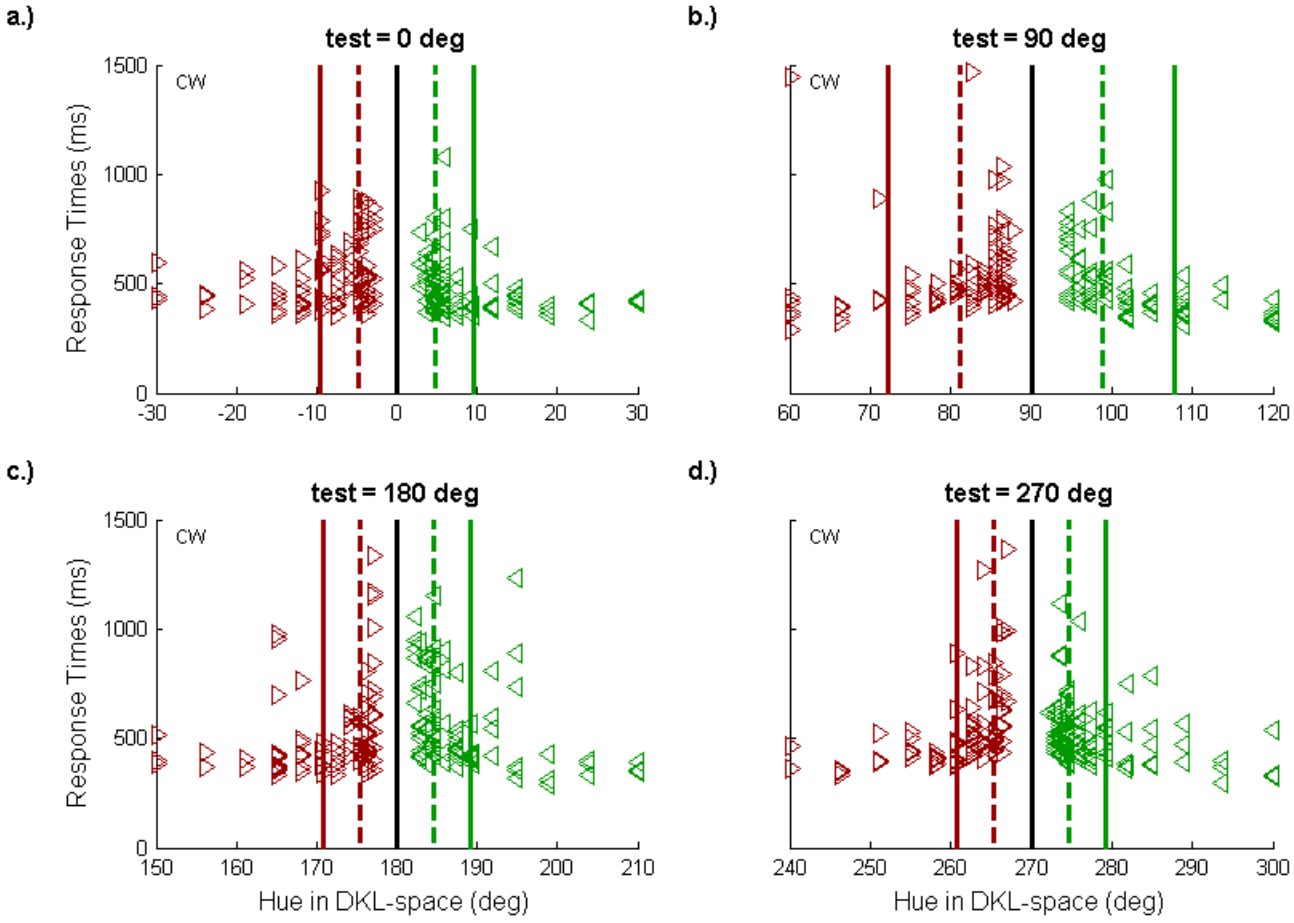

Figure S1. Response times at threshold. The x-axis represents hue along the DKL-circle in azimuth degree, the y-axis response times of observer CW in the preliminary JND measurements. Each of the four panels show the response times at an example test color, which is at 0 degree (panel **a**), 90 degree (**b**), 180 degree (**c**), and 270 degree (**d**). Each triangle refers to a trial of an azimuth-increasing (red) or azimuth-decreasing (green) staircase. The vertical black line indicates the hue of the test color, the colored dashed lines 1 JND, and the colored solid lines 2 JNDs for either hue direction. *Note that beyond 2 JNDs response times are much less variable and are about stimulus offset time (500ms) for all test colors. These observations are discussed in the sections "Stimuli", "Response times and JND measurements", and "Task sequence and task demands" of the* main article.

Individual differences of stimulus pairs (Fig. S2)

Figure S2 complements Figure 3.c in the main article. These figures illustrate individual differences in categorization, discrimination, and equally discriminable color pairs (for further details on individual differences see Witzel & Gegenfurtner, 2013, p. 6-15). Figure 3.c in the main article shows the individual categories in DKL space. Panel a of this Figure S2 illustrates individual differences in the sensitivity to color differences in DKL space. In this graphic, DKL-space is divided into 10 parts that contain an equal number of JNDs. Panel c shows categories in relative JND-space (for explanations of relative JND space see section "Main Results: Category effects" in the main article). Due to individual differences in the total number of JNDs, absolute JND space would result in different lengths of the axis of JND space for each individual. Relative JND space allows to align the axes of each individual for illustration purposes.

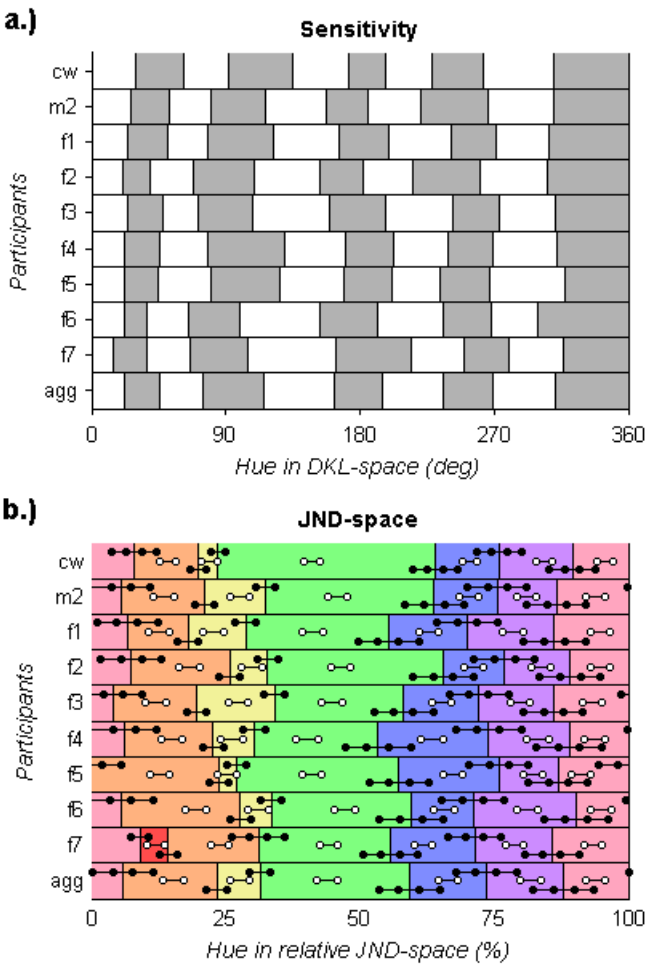

Figure S2. Interindividual differences in sensitivity and stimulus pairs. In both panels, rows along the y-axis correspond to participants, and the last row to the aggregated data (agg), as in the corresponding Figure 3.c of the main article. Panel **a** illustrates the variation of sensitivity in DKL-space. The x-axis represents hue in azimuth degree as in Figure 3.c. The isoluminant circle of DKL-space is divided in ten sections that are equally large in terms of JNDs. These sections are represented as alternating grey and white areas. Panel **b** illustrates categories and stimulus pairs in relative JND-space. The x-axis corresponds to hue in relative JND-space, specified in percent of the overall hue circle. Apart from that format as in Figure 3.a & c. *Note the differences between the individual observers*. *The way in which categories differ between* Figure 3.c *and panel b of this figure depends on the variation of sensitivity illustrated in panel a. While equally discriminable color pairs for each observer have different sizes in DKL-space, they have the same size in relative JND-space (panel b). Across observers the size of the stimulus pairs slightly differs in relative JND-space depending on each observers overall sensitivity (panel b).*

Colorimetric specifications of stimuli

Chromaticity diagram (Fig. S3)

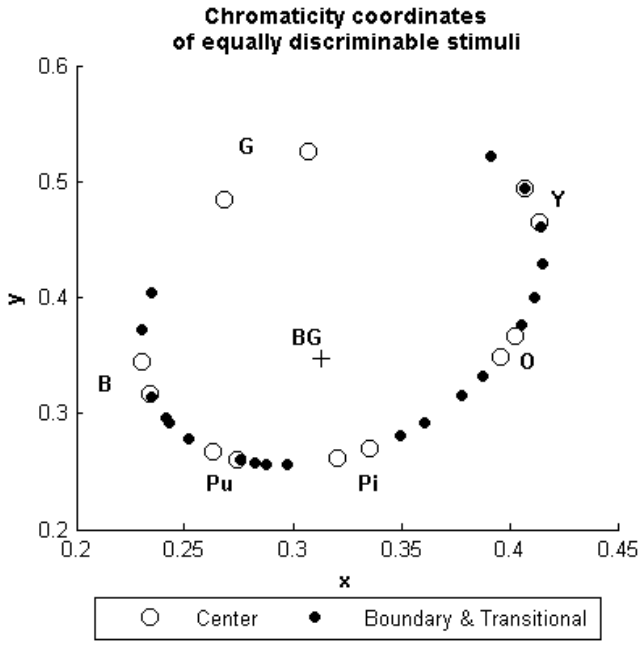

Figure S3. Chromaticity diagram of equally discriminable stimuli. X- and y-axes refer to the chromaticity coordinates x and y, respectively. The + refers to the chromaticities of the grey background ("BG"), the white disks to those of the stimuli of the center pairs ("Center"), and the black disks to those of the boundary and transitional pairs ("Boundary & Transitional"). The letters close to the coordinates of the center pairs correspond to the initials of the respective color names (e.g. G = Green; cf. Figure 3.a). *The coordinates correspond to those given in Table S1.*

Chromaticity coordinates and luminance (Tab. S1)

Category	Symbol	Pair type	Color 1			Color 2		
			azi	x	y	azi	x	y
Background (BG)	-	-	-	0.3127	0.3476	-	-	-
Pink (Pi)	●	Pink-orange transitional	0.3	0.3783	0.3151	9.3	0.3879	0.3318
	●	Pink-orange boundary	9.3	0.3879	0.3318	17.7	0.3961	0.3491
Orange (O)	●	Orange-pink transitional	17.7	0.3961	0.3491	25.5	0.4029	0.3671
	○	Orange center	29.3	0.4059	0.3765	38.2	0.4114	0.3995
	●	Orange-yellow boundary	49.0	0.4151	0.4290	60.3	0.4145	0.4609
Yellow (Y)	○	Yellow center	61.6	0.4141	0.4644	72.9	0.4071	0.4944
	●	Yellow-green boundary	73.0	0.4071	0.4945	85.9	0.3913	0.5223
Green (G)	○	Green center	125.7	0.3071	0.5263	144.8	0.2685	0.4850
	●	Green-blue transitional	174.3	0.2349	0.4035	186.6	0.2303	0.3717
	●	Green-blue boundary	186.6	0.2303	0.3717	198.4	0.2301	0.3447
Blue (B)	●	Blue-green transitional	198.4	0.2301	0.3447	214.0	0.2347	0.3147
	○	Blue center	212.4	0.2340	0.3174	228.6	0.2427	0.2927
	●	Blue-purple transitional	226.1	0.2411	0.2961	241.4	0.2519	0.2780
	●	Blue-purple boundary	241.4	0.2519	0.2780	254.6	0.2631	0.2669
Purple (Pu)	●	Purple-blue transitional	254.6	0.2631	0.2669	266.7	0.2743	0.2602
	○	Purple center	268.2	0.2757	0.2596	280.2	0.2879	0.2564
	●	Purple-pink transitional	275.1	0.2826	0.2574	289.4	0.2975	0.2560
	●	Purple-pink boundary	289.4	0.2975	0.2560	310.0	0.3203	0.2614
Pink (Pi)	●	Pink-purple transitional	310.0	0.3203	0.2614	335.0	0.3493	0.2804
	○	Pink center	323.4	0.3358	0.2698	344.9	0.3608	0.2920

Table S1. Chromaticitiy coordinates of the aggregated equally discriminable stimuli for the second group. "Category" refers to the category membership of the stimuli, "Symbol" to the symbol in Figure 3 and Figure S2, "Pair type" to the type of color pair, "azi" to azimuth in DKL-space in degree, "x" and "y" to (computed) Judd-corrected chromaticity coordinates. Luminance of all colors was computed as 30.05 cd/m^2, and about 28 cd/m^2 when measured on the screen. The chromaticity coordinates and luminance of the background were x = 0.3129, y = 0.3484, Y = 27.9 cd/m^2 when measured on the screen (for further details see Witzel & Gegenfurtner, 2013). *Note that colors were fitted into the gamut of the monitor used in this study, and might be outside the gamut for other monitors and settings.*

Feedback and Hall of Fame

In order to motivate participants to maximise speed and accuracy we provided qualitative feedback after each block, and at the end of all blocks, they could enter their name in a hall of fame depending on their performance (cf. section "Procedure" of the main article). For both, block-wise feedback and hall of fame, a score has been calculated based on the product of the normalized response times and accuracy rates and mapped to a scale between 0 and 10000. So, accuracy and speed could compensate each other and the same score could be obtained with a relatively low accuracy rate and a respectively higher speed or vice versa. 10000 points corresponded to 100% correct and an average response time of 500ms. Hence, with response times lower than 500ms participants could obtain more than 10000 points. 0 points corresponded to a performance of 50% correct answers or an average RT of more than 1500ms.

Feedback. The block-wise feedback was given by written ratings on the screen. A performance better than 10000 points was called "Fantastic!!!"; for performance corresponding to more than 95% correct and response times of less than 600ms (8100 points), participants obtained "Excellent!!", for more than 90% correct and less than 700ms "Very Good!", for 85% and 700ms "Good" and for over 50% and less than 1000ms only "Ok". If participants yielded less than 50% correct or more than 1500ms (0 points) they were asked to contact the experimenter.

Hall of fame. The hall of fame consisted of 10 entries, ordered by the scores. Participants could enter their name at the respective rank if they obtained a higher score than at least one of the listed scores. At the beginning of data collection, entries were set to 0 so that in the first ten sessions all people entered the hall of fame.

MAIN RESULTS

Categorical perception tests

Group 1 (Tab. S2)

Border	Response Times (ms)				Error rates (%)			
	M	*df*	*t*	*p*	*M*	*df*	*t*	*p*
a.) transitional - boundary								
Pi-Pu	-4.5	8	-0.4	0.74	-0.6	8	-0.5	0.62
Pi-O	29.2	7	1.3	0.22	3.3	7	1.6	0.15
O-Pi	-4.5	7	-0.2	0.82	-0.3	7	-0.1	0.90
G-B	25.2	8	1.8	0.10	0	8	0.1	0.99
B-G	-32.6	8	-3.4	*	-4.3	8	-2.3	*
B-Pu	6.0	8	0.4	0.69	-2.2	8	-1.2	0.28
Pu-B	20.5	8	2.1	°	2.3	8	1.3	0.24
Pu-Pi	-6.4	8	-0.4	0.72	-0.7	8	-0.4	0.71
b.) center - transitional								
Pi-Pu	34.5	8	1.8	0.1	1.5	8	1.5	0.17
Pi-O	10.9	7	0.3	0.76	-1.0	7	-0.5	0.65
O-Pi	11.5	7	0.6	0.55	1.2	7	1.0	0.36
G-B	38.6	8	2.2	°	6.3	8	4.4	**
B-G	29.0	8	3.5	**	0.2	8	0.2	0.86
B-Pu	-9.5	8	-0.5	0.61	-2.0	8	-1.5	0.18
Pu-B	-18.2	8	-1.0	0.35	-0.8	8	-0.3	0.80
Pu-Pi	8.7	8	0.7	0.50	2.2	8	0.8	0.47

Table S2. Categorical perception tests for transitional pairs in group 1. Part **a** reports t-tests for comparisons between transitional and boundary pairs; part **b** for transitional and center pairs. Format as in Table 1 and Table 2 of the main article.

Group 2 (Tab. S3)

Border	Response Times (ms)				Error rates (%)			
	M	*df*	*t*	*p*	*M*	*df*	*t*	*p*
a.) transitional - boundary								
Pi-Pu	41.9	11	3.5	**	3.1	11	3.2	**
Pi-O	98.9	11	6.7	***	3.3	11	3.3	**
O-Pi	64.4	11	4.4	**	1.7	11	1.9	°
G-B	39.6	11	3.3	**	2.5	11	3.2	**
B-G	-86.5	11	-7.0	***	-3.3	11	-5.5	***
B-Pu	8.5	11	1.1	0.30	0.43	11	0.5	0.65
Pu-B	29.8	11	2.4	*	-0.1	11	-0.1	0.89
Pu-Pi	32.7	11	2.7	*	2.7	11	2.5	*
b.) center - transitional								
Pi-Pu	78.7	11	5.0	***	3.0	11	1.7	0.12
Pi-O	21.8	11	1.4	0.19	2.8	11	2.6	*
O-Pi	44.1	11	4.5	***	2.1	11	1.6	0.14
G-B	9.5	11	0.4	0.70	4.2	11	3.0	**
B-G	25.2	11	4.4	**	0.4	11	0.7	0.48
B-Pu	-69.8	11	-5.4	***	-3.3	11	-2.7	*
Pu-B	12.4	11	0.9	0.40	0.8	11	0.6	0.55
Pu-Pi	9.6	11	0.8	0.46	-2.0	11	-2.2	°

Table S3. Categorical perception tests for transitional pairs in group 2. Format as in Table S2.

Independence of categorical patterns (Tab.S4)

The binomial tests reported in the section "Group 2 (initially inexperienced observers)" of the main article require statistical independence of random events. To get an idea of the statistical dependencies of the observed categorical patterns in the second group of participants, we calculated the correlations across the 12 participants of the relative response times of the center pairs for the 6 categories (left side of Table S4), and those for the relative error rates (right side of Table S4). None of the relative response times were correlated between any of the categories, and only the relative error rates of the pink and orange categories were significantly correlated, explaining 36% of the variance ($r = 0.60$, $p = 0.04$).

Note that correlations may also result from relationships of actual category effects across participants, rather than from correlations of error terms. However, the general absence of correlations indicates that neither category effects, nor error terms are correlated.

Clearly, correlations only capture linear relationships, and might miss other statistical dependencies. Nevertheless, the absence of strong correlations indicates that there were no obvious statistical dependencies across the categories. For this reason, we used the binomial distribution as an approximation for estimating the probabilities that the categorical patterns observed in the second group occurred by chance.

At the same time, the absence of correlations indicates that there were no systematic variations of category effects across individuals. This observation is also relevant for the differences in categorical patterns between the two groups, and is discussed for this reason in the section "Individual differences" of the main article, and in section "Individual Observers (Fig. S4)" of the Supplementary material.

		Pink		Orange		Yellow		Green		Blue		Purple	
	n	r	p	r	p	r	p	r	p	r	p	r	p
Pink	12	1	0	0.60	*	0.49	0.10	0.20	0.53	-0.27	0.40	-0.29	0.37
Orange	12	0.39	0.22	1	0	0.23	0.47	0.07	0.84	-0.03	0.92	0.20	0.52
Yellow	12	0.01	0.97	0.02	0.96	1	0	0.30	0.35	-0.25	0.44	-0.49	0.10
Green	12	0.17	0.59	0.10	0.77	0.49	0.11	1	0	0.01	0.98	0.17	0.59
Blue	12	0.10	0.76	0.24	0.45	0.37	0.24	0.37	0.23	1	0	-0.06	0.85
Purple	12	-0.12	0.71	-0.11	0.74	0.48	0.11	0.20	0.53	-0.07	0.83	1	0

Table S4. Correlations of relative response times and error rates between center pairs of different categories in the second group. Correlations between relative response times are shown in green (left side), those between relative error rates in red (right).

ADDITIONAL ANALYSES OF MAIN RESULTS

Individual observers (Fig. S4)

We examined whether category effects only occurred for some observers, but not for others. For this purpose, we calculated the mean response times and error rates for each block across sessions (15 blocks) for each individual observer. We applied the same tests as for the aggregated data across individuals to the individual data across blocks. Results for the centre pairs are illustrated in Figure S4. In the case of a category effect, bars should be above zero.

The left part of Figure S4 shows the results for the first group. Apart from the green and the blue category, results do not show a consistent pattern, neither across categories nor across individuals. In fact, in those categories (pink, orange, yellow, and purple) some bars went significantly in the direction of the category effect, and others in the opposite direction. This was true for response times (panel a) as well as for error rates (panel b). However, the pattern of response times (bars in panel a) and the pattern of error rates (bars in panel b) across observers and categories were correlated ($n = 54$; $r = 0.60$; $p < 0.001$). This was also true for the distances of the transitional pairs from the boundary line ($n = 72$, $r = 0.66$, $p < 0.001$), and the differences between these distances for center and transitional pairs ($n = 63$, $r = 0.58$, $p < 0.001$). These correlations suggest that the idiosyncratic variations of response times and error rates reflect systematic inter-individual differences in performance across color pairs.

Panels c and d of Figure S4 illustrate the individual results for the centre pairs of the second group. For pink, orange, yellow, and green, relative response times and error rates were above zero for almost all individuals, which is in line with the category effect. However, blue consistently shows the inverse pattern. Only purple indicates strong differences across individuals. As for group 1, response times and error rates were correlated; but correlations were lower than those for group 1 (centre pairs: $n = 72$, $r = 0.35$, $p = 0.002$; transitional pairs: $n = 96$, $r = 0.32$, $p = 0.002$; centre vs. transitional: $n = 96$, $r = 0.19$, $p = 0.06$).

Moreover, we reported in section "Independence of categorical patterns" that categorical patterns were not correlated across participants in the second group (Table S4). If some observers were more susceptible to category effects than others, the strength of categorical patterns should vary consistently for all categories. Hence, categorical patterns in different categories should be correlated across observers. The absence of such a correlation suggests that category effects do not systematically vary across observers. This undermines the idea that differences in categorical patterns between groups could be due to differences in the susceptibility for category effects across individuals.

As summarized in the main article (section "Individual observers), facilitation effects were consistent across observers in the second group (Figure S4.c-d). In the first group, only green and blue yielded consistent effects across observers, green in line with categorical facilitation, blue in the opposite direction. There were no individuals in the first group that showed consistent effects across categories (Figure S4.a-b). These results further indicate that there were fundamental differences between the two groups.

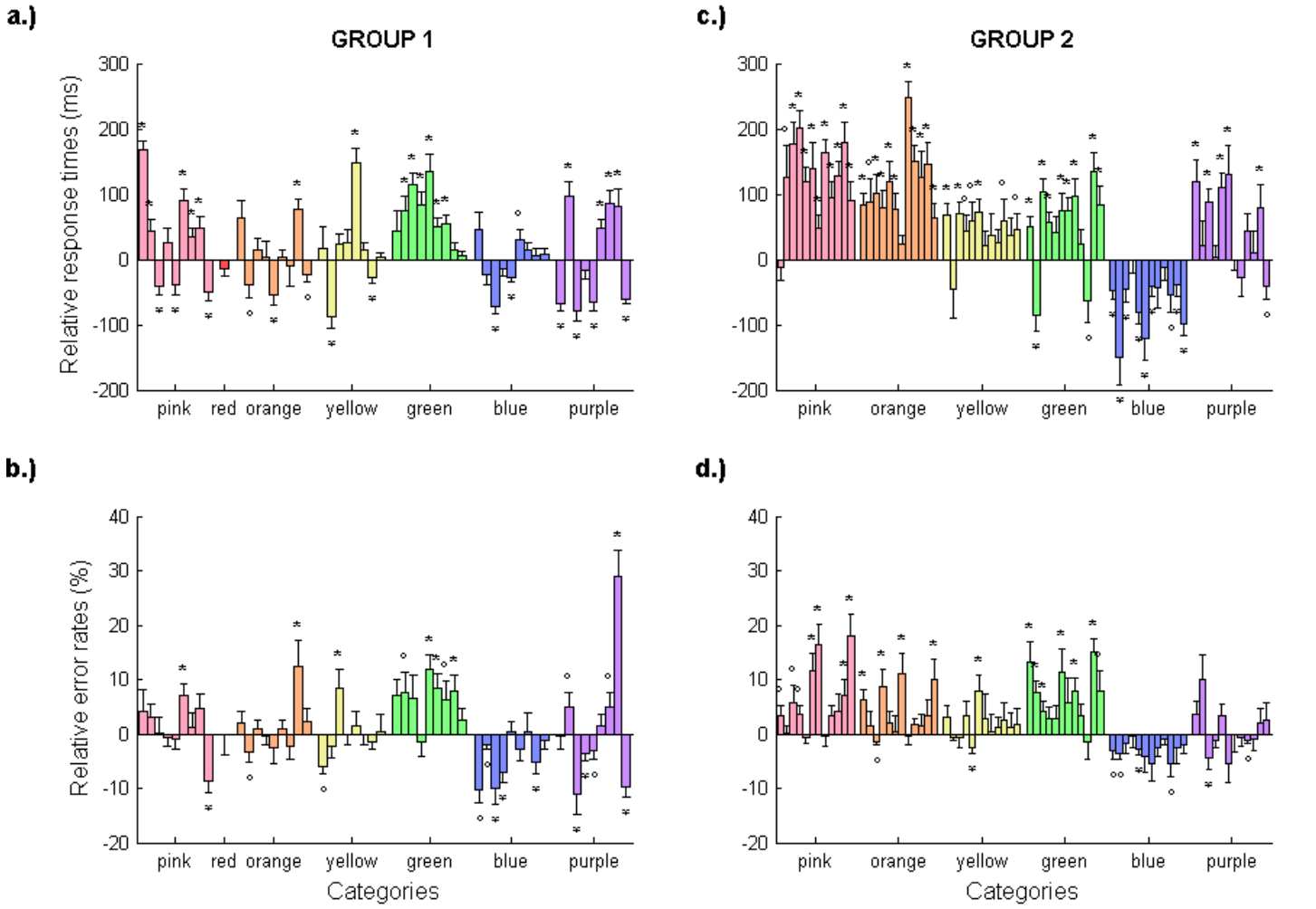

Figure S4. Individual category effects. Average relative response times (panels **a & c**) and error rates (**b & d**) for the center pairs are shown for the individual observers of the first (**a & b**) and second group (**c & d**). The bars along the x-axis correspond to the average of the single observers, ordered by categories. The colors of the bars and the labels along the x-axis refer to the respective color terms. Error bars correspond to standard errors of mean across 15 blocks. The symbols above the bars refer to the p-values of paired, two-tailed t-tests, testing for the difference from zero across blocks, with * corresponding to $p < 0.05$ and ° to $p < 0.1$. *Note the high variability of the direction of the bars for all categories except green in the first group, and the almost unanimous orientations of the bars within each category in the second group.*

Response Time Distributions

Cumulative density functions (Fig. S5-S6)

The distribution of response times may be illustrated by cumulative density functions of response times. Cumulative density functions show the number of responses with response times below or equal to the response times indicated along the x-axis (These distributions are sometimes also called *vincentized*). We plotted these cumulative density functions lumped together across participants, but separately for each category (cf. different panels) and separately for each type of stimulus pair (red for center vs black curves for boundary pairs). The higher the curves in these graphics, the faster were the response times. According to a category effect, the red curve for the center pairs should be lower than the two black curves for the boundary pairs.

Figure S5 shows the cumulative density functions for group 1. For green (panel d) the curve for the center pair (red) is lower than those of the boundary pairs in the last quartile (above 75%), which is in line with the category effect. For the other categories, these curves almost completely coincided for all quantiles, indicating that there is really no category effect in the aggregated data.

In sum, no category effects appeared in the first group except for the green category (cf. section "Response Time distributions of the main article).

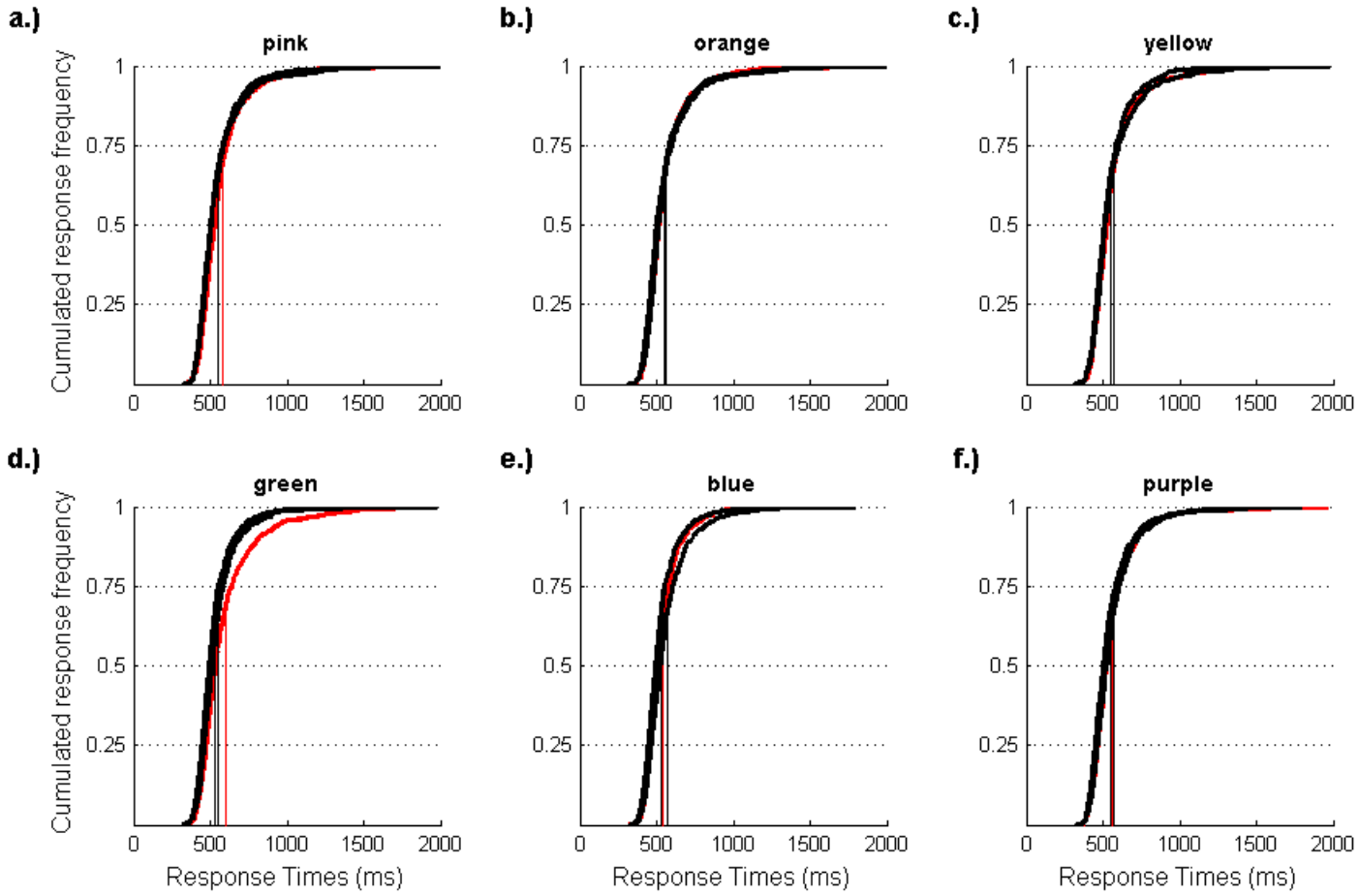

Figure S5. Cumulative density function of response times for group 1. Response times of all individuals were lumped together, but separated per category. The x-axis refers to response time margins, and the y-axis to the relative frequency of correct responses that were given faster or equal to the margins. Red curves correspond to the response times for center pairs, black curves to those of the boundary pairs. Horizontal dotted lines indicate quartiles, while averages are shown by the vertical lines. *Note that the red curve of the center pair is lower than the black ones of the boundary pairs for the green, but for none of the other categories.*

Figure S6 shows the cumulative density functions for group 2. In line with a category effect, the red curve for the center pairs is lowest for pink (panel a), orange (b), yellow (c), green (d), and purple (f). These are the same categories that showed the categorical patterns with average response times (Figure 4.c-d). The maximum differences between the curves are located above the average, at about the 75th percentile. This indicates that categorical patterns occur for response times at about this percentile rather than around the median or below.

Taken together, the results for both groups are in line with the main results for the average response times in Figure 4 of the main article (cf. section "Response Time distributions of the main article).

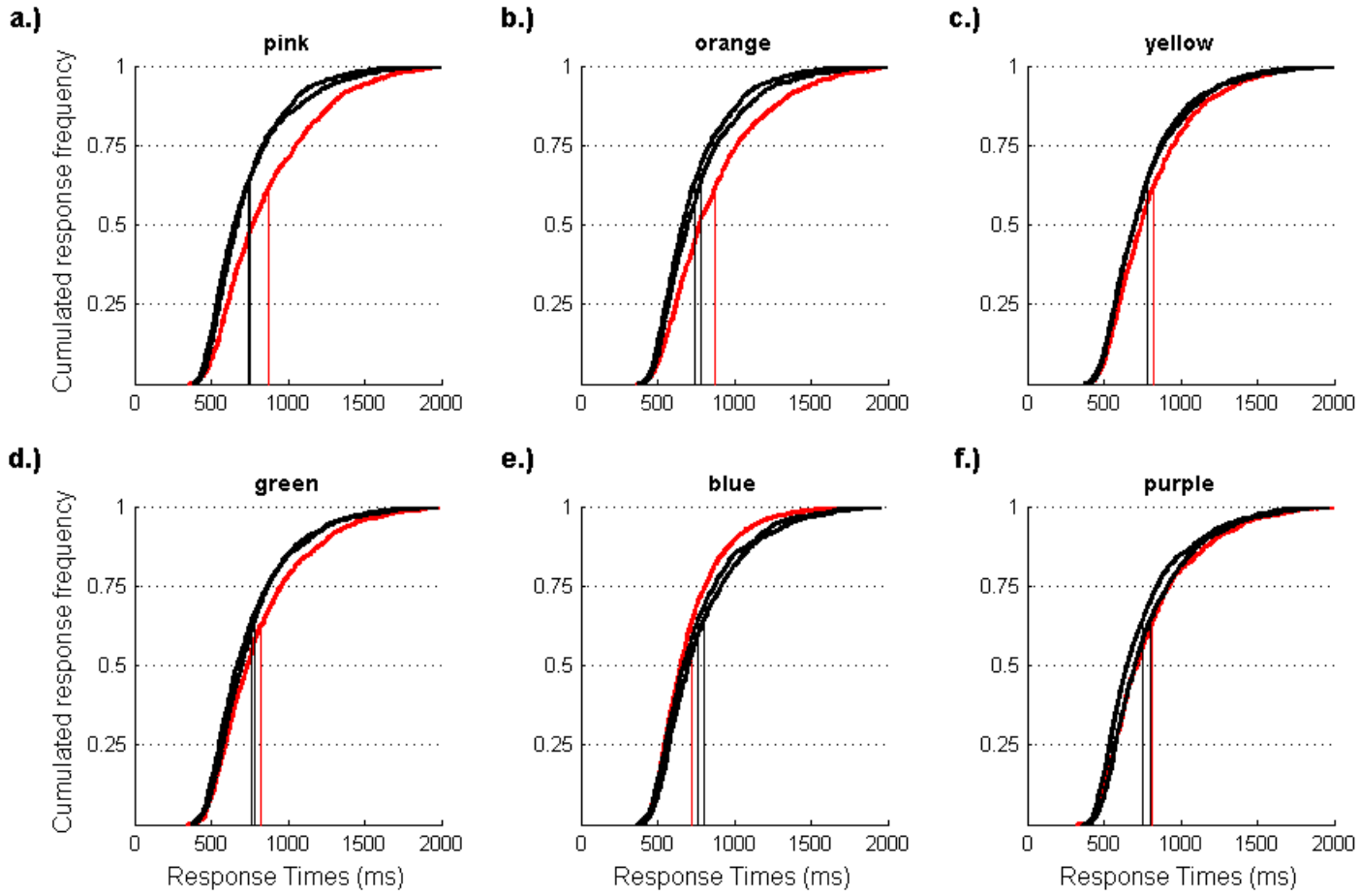

Figure S6. Cumulative density functions of response times for group 2. Format as in Figure S5. Above the *median the red curves of the center pairs are lower than the black curves of the boundary pairs for all categories except blue, for which this pattern is reversed.*

Percentiles and cut-offs (Fig. S7)

Data at different levels of response speed were analyzed to statistically test the dependence of category effects on the distribution of response times. In order to inspect performance relative to each observer's individual response speed, we divided each participant's data into ten parts based on response time deciles (*relative partitions*). To examine absolute cut-off values of response times, we also separately analyzed response times below and above 700ms (*absolute partitions*). This cut-off value is (approximately) the median of the second group and still allows for sufficient data for each participant in the first group (cf. section "Overall Performance" in the main article). Paired, two-tailed t-tests across participants were applied to the average relative response times and error rates of the respective partitions of the data. Results for the center pairs are illustrated by Figure S7. Light, unsaturated bars refer to the relative partitions based on deciles, dark, saturated bars refer to the absolute partition at 700ms.

The second, initially inexperienced group, showed clear patterns across the 10 relative partitions (Figure S7.c-d). The 5 categories that yielded category effects with the overall averages (pink, orange, yellow, green, and purple) also yielded such effects for the 10 relative partitions (cf. light, unsaturated bars in Figure S7.c-d). Relative response times of the center pairs were higher than zero for these categories (panel c). The same was true for most of the relative error rates; though the effects on error rates were less consistent, in particular for yellow and purple (panel d). For blue, the relative response times and error rates were below zero for almost all relative partitions. The size of the difference between the response times for the center pair and the boundary line systematically increased with the absolute size of the response times, as illustrated by the increasing height of the bars in panel c.

The absolute cut-off did not yield as clear patterns as the deciles (dark, saturated bars in Figure S7.c-d). For response times above 700ms, the relative response times of the center pair were significantly different from zero for all categories, except yellow and purple, where differences were marginally significant and non-significant, respectively. For response times below 700ms only orange yielded a significant difference from the boundary line. At the same time, error rates yielded similar category effects below and above the 700ms cut-off (dark, saturated bars in Figure S7.d).

In the first group, green yielded a similar pattern of relative response times and error rates across deciles as in the second group (cf. light bars in Figure S7.a-b). In contrast, the inverse effect for blue response times occurs mainly for response times above 700ms. Finally, the error rates for orange show categorical patterns when partitioned into deciles or when partitioned according to the 700ms cut-off. These categorical effects did not appear for the overall average (cf. Figure 4.b of the main article). Apart from that, no additional category effects were found in the first group.

In sum, apart from the additional effects for orange in the first group, these results mainly confirm the main results (cf. Figure 4 of the main article) and show their robustness across the response time distribution. The green category in the first group and the 5 categories in the second group yielded category effects across all 10 response time deciles of each participant, and for response times below and above an absolute cut-off of 700ms. Hence, observed categorical patterns neither depend on the size of response times relative to each observer, nor on the absolute size of response times (cf. section "Response time distributions" in the main article).

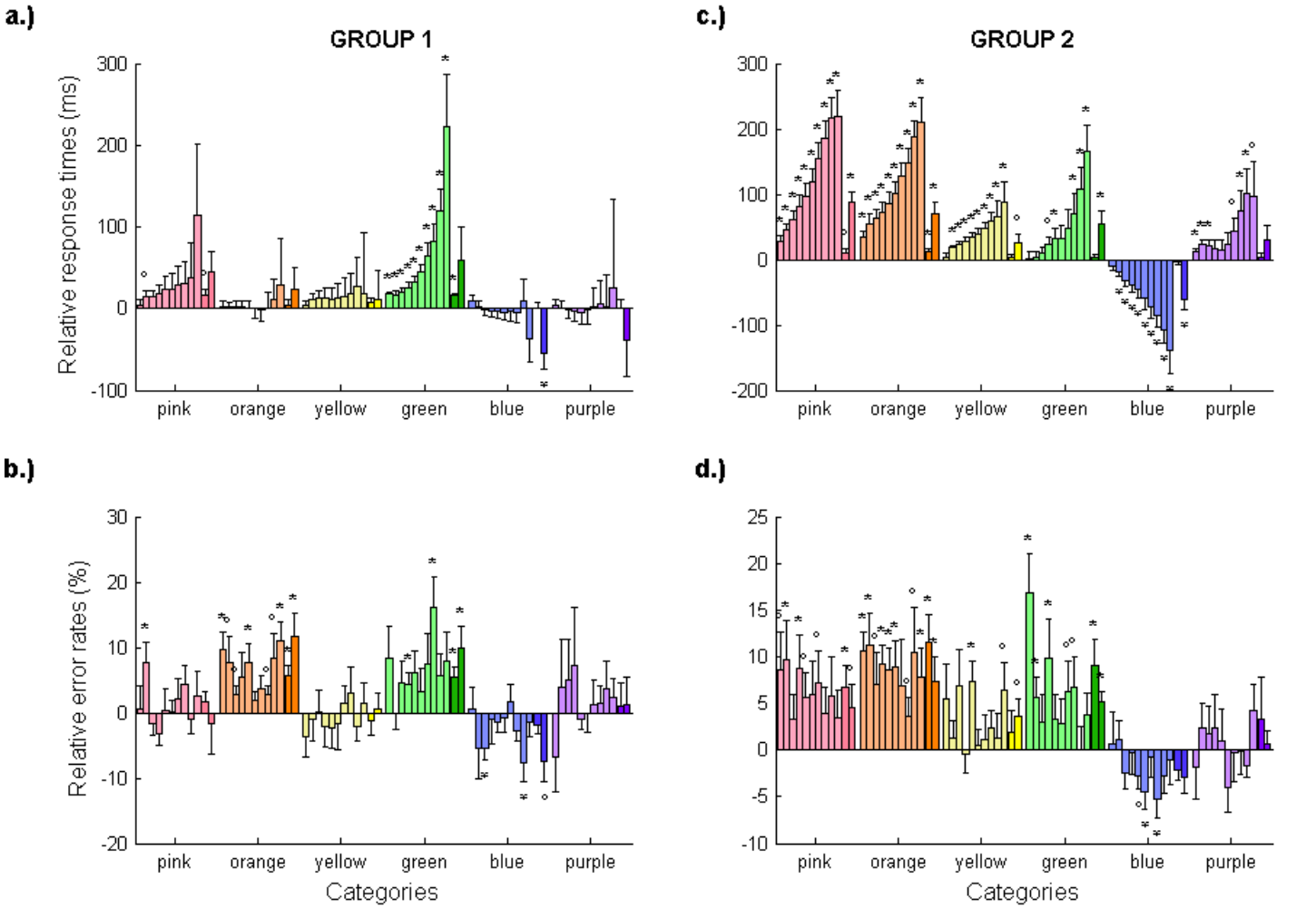

Figure S7. Category effects across percentiles and 700ms cut-off. The y-axis represents relative response times (panels **a & c**) and error rates (**b & d**) of the center pairs in the first (**a & b**) and second group (**c & d**). Different bars refer to different partitions of data. Groups of bars correspond to categories, as identified by the bar colors and the labels along the x-axis. Light, unsaturated bars (the first 10 in each group) show averages for data that has been divided in 10 parts by the response time deciles of each participant separately (relative partitions). The dark, saturated bars (last two in each group) refer to data, for which response times were below and above 700ms (absolute partitions). Error bars correspond to standard errors of mean across participants. Symbols refer to p-values of paired, two-tailed t-tests across participants, with * corresponding to $p < 0.05$ and ° to $p < 0.1$. *Note that the category effects for pink, orange, yellow, green, and purple in the second group (panels c and d) and for green in the first group (panels a and b) consistently occur in all partitions.*

Time and Training

Performance over time (Fig. S8)

Figure S8 shows the development of performance across the overall 15 blocks (3 blocks x 5 sessions). Average response times and error rates were calculated separately for center (white curve), boundary (black), and the two transitional pairs (green and red).

In the first, highly trained group (left column of Figure S8), response times increased up to the fourth session, and decreased towards the 5th session (Figure S8.a). Error rates showed the inverse pattern across blocks (Figure S8.b). As a result, response times and error rates were negatively correlated (min r = -0.65, p < 0.01), indicating a speed-accuracy trade-off.

In contrast, in the second group response times for all 4 stimulus types decreased with time. Consequently, they were significantly correlated with block order (minimum r = -0.82, all p < 0.001). However, error rates did not change systematically across blocks (max r = 0.14, p = 0.62). Hence, there was no speed accuracy trade-off. The increase in speed must be attributed to learning.

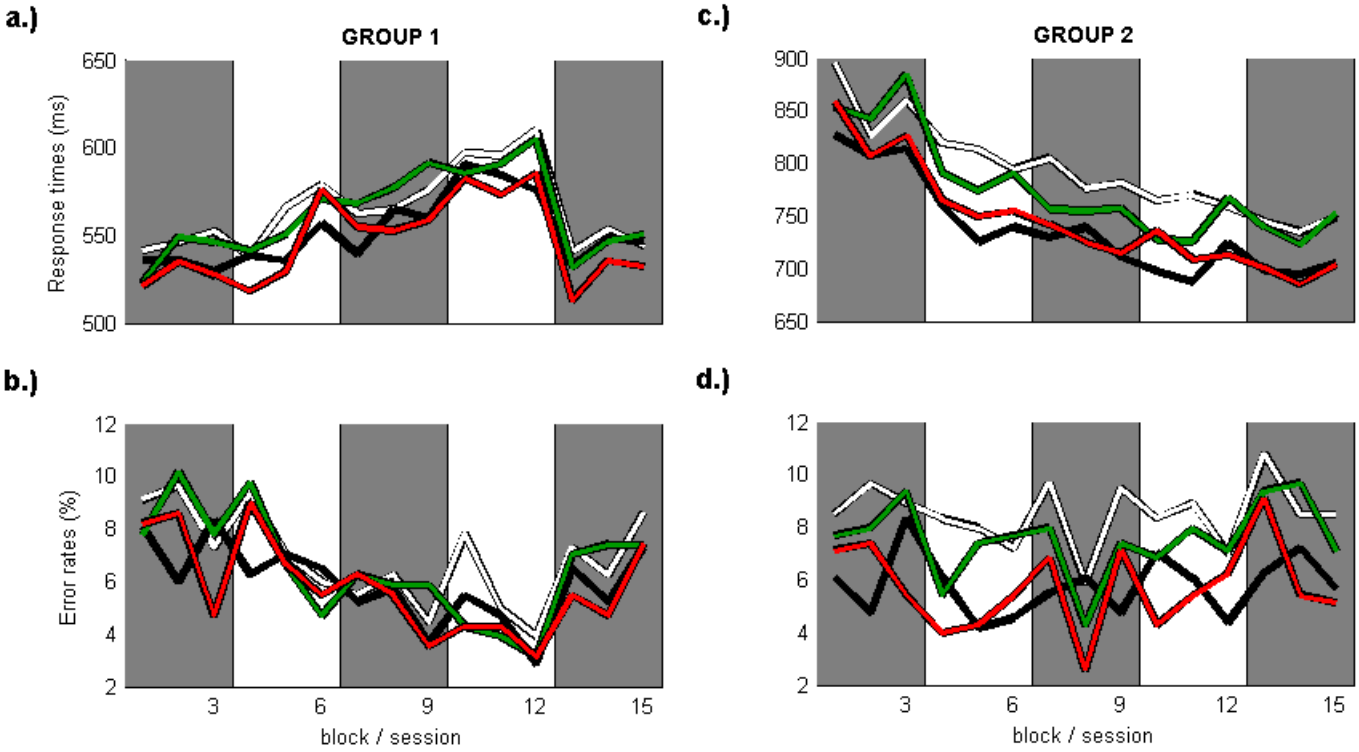

Figure S8. Performance across blocks. The y-axis represents average response times (panels **a & c**) and error rates (**b & d**) for the different stimulus types in the first (**a & b**) and second group (**c & d**). The x-axis corresponds to the 15 blocks, the alternating grey and white background illustrates the 5 sessions. Data is shown separately for center (white curve), boundary (black curve), and the two kinds (lower and upper azimuth) of transitional pairs (green and red curves). *Note that response times steadily decrease across blocks in the second but not in the first group, while there is a speed-accuracy trade-off in the first but not in the second group.*

Feedback Scores (Fig. S9)

We tested whether the reinforcement scheme used for feedback and hall-of-fame affected the development of performance. In particular, it may have provided reinforcement patterns that weigh response times and error rates differently at different levels of response speed. The first group might optimize speed-accuracy trade-offs according to the particular reinforcement scheme. Figure S9 illustrates the development of scores and feedback across blocks.

In the first group (Figure S9.a), only 2 out of 9 participants (f3 & f4) consistently improved their scores, as shown by a significant positive correlation between scores and blocks (r(15) = 0.59, p = 0.02; and r(15) = 0.77, p < 0.001; all other p > 0.11). Note that participants did not know the exact scores, but just the qualitative feedback. The two participants that improved their scores (purple and yellow line) changed from "very good" during the first session, to "excellent" in later sessions.

In the second group (Figure S9.b), 6 out of 12 participants significantly improved their scores across blocks (r > 0.57, all p < 0.03. Another two participants improved marginally significantly (r = 0.50, p = 0.06, and r = 0.48, p = 0.07; all other p > 0.13). These participants mainly improved from "good" in the first, to "very good" in later ses-

sions. However, none of the participants in the second group ever reached the level for "excellent".

These results show that the idiosyncratic patterns of response times and error rates across blocks in the first group did not serve the purpose of achieving higher scores or better feedback.

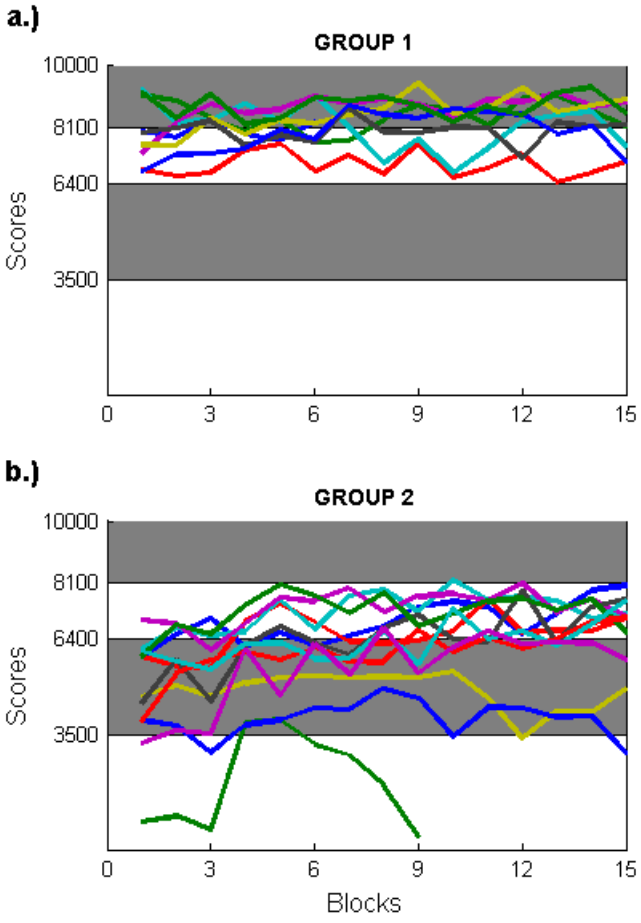

Figure S9. Feedback scores across sessions. The x-axis corresponds to the 15 blocks across 5 sessions. The y-axis represents the scores that were calculated based on response times and error rates at the end of each block. Block-wise feedback was "fantastic" for scores above 10.000, "excellent" > 8100, "very good" > 6400, "good" > 3500, and "ok" below 3500. Each curve represents one participant of the first group (panel **a**) or the second group (panel **b**). *Note that few participants of the first group, but several of the second group consistently improved their scores and feedback across blocks.*

Category effects over time (Fig. S10)

To assess whether category effects disappear with training and experience, we examined the development of category effects over time. For this purpose, we concentrated on the center pairs, and we calculated relative response times and error rates (distances to the boundary lines) separately for each block. Figure S10 shows the relative response times (first row) and error rates (second row) of the second group. The curves correspond to the median across participants. If category effects decreased with familiarity and training, curves should decrease across blocks.

To test whether category effects changed systematically across blocks, correlations were calculated between relative block and relative response times and error rates. To test for category effects within blocks, paired, two-tailed t-tests across participants were used.

Figure 10.a-b illustrates the results for the first group. If training and experience would have counteracted category effects in the first group, category effects would have existed at the beginning and faded across blocks and sessions. However, this was not the case. Apart from green and blue, only the relative response times for purple were marginally significant above zero in the second block ($M = -37.5$, $t(8) = -2.0$, $p = 0.08$). Neither relative response times nor error rates of any other category showed a significant difference in any of the three blocks of the first session (min. $p = 0.13$). Moreover, none of the relative response times and error rates were negatively correlated with blocks, apart from a significant negative correlation of purple error rates ($r(15) = -0.59$; $p = 0.02$), and a marginally significant negative correlation for pink response times ($r(15) = -0.49$, $p = 0.06$).

As reported above ("Performance over time"), the second group showed increasing performance across blocks, indicating an effects of training and experience on overall performance (Figure S8.c-d and Figure S9.b). If training and experience affected the category effects of the second group, category effects should attenuate across blocks. Yet, this was not the case (Figure S10.c-d). Category effects existed across all blocks in the 5 categories, pink, orange, yellow, green, and purple. The opposite pattern in the blue category was also stable across blocks. Moreover, the category effects in the second group did not systematically decrease over time. Only the relative response times for pink were negatively correlated with blocks ($r(15) = -0.55$, $p = 0.03$). For green relative response times ($r(15) = 0.52$, $p = 0.049$) and error rates ($r(15) = 0.56$, $p = 0.03$) were positively correlated with blocks. These positive correlations imply that the categorical patterns in these two categories becomes more pronounced over time. Blue also showed a positive correlation, but only for response times ($r(15) = 0.60$, $p = 0.02$). This correlation indicates that the inverse pattern in the blue category decreases over time.

In sum, the correlations did not show a consistent decrease of category effects with time and training. The conditions (groups and categories) that yielded category effects did so across all sessions, and no additional category effects appeared in the first group when focussing on single blocks. Hence, there is no evidence for a modulation of category effects across blocks. These results undermine the idea that categorical facilitation is affected by training and experience with the task.

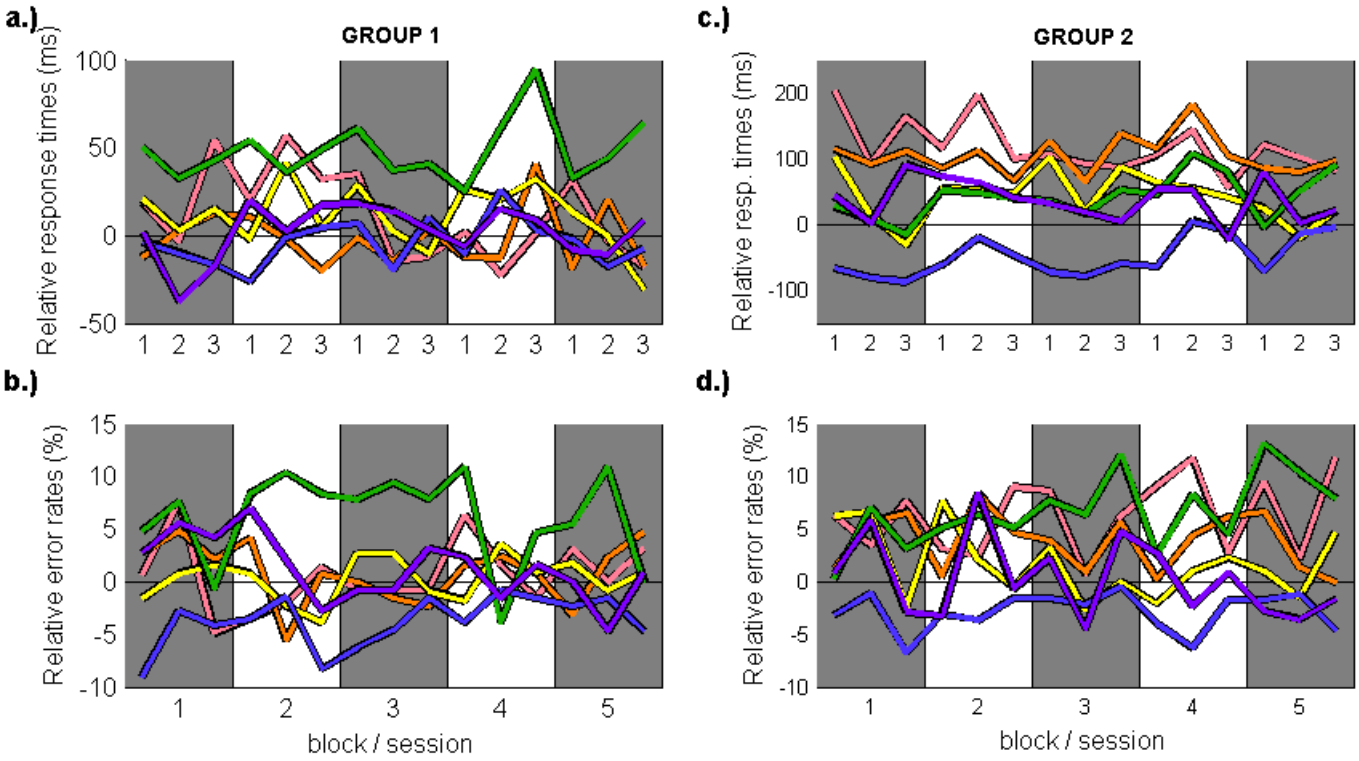

Figure S10. Category effects across blocks. The left side corresponds to results of the first group (panels **a & b**), the right side to those of the second group (**c & d**). The upper rower shows response times (**a & c**), the lower row error rates (**b & d**). The x-axis represents blocks (grey and white backgrounds refer to session). The y-axis corresponds to the relative response times or error rates per block, respectively. The different curves refer to the center pairs of the different categories; colors of curves indicate categories. *Note that the second group shows similar patterns in each block, with little modulation over time (panels c and d). In contrast, there are almost no stable patterns over time in the first group (panels a and b). In the first group, only the categorical pattern in the green category is stable across blocks, and the relative error rates in the blue category are negative across all blocks.*

T-Tests across blocks

To achieve higher statistical power than in the main tests for category effects across participants, we also conducted t-tests for category effects across the 15 blocks, with data aggregated across participants.

In the first group (Figure S10.a-b), average response times and error rates for green were above the boundary line in almost all blocks. Hence, t-tests across blocks were significant for response times ($t(14) = 10.9$; $p < 0.001$) and error rates ($t(14) = 5.3$; $p < 0.001$). Response times for pink ($t(14) = 2.0$; $p = 0.07$) and yellow ($t(14) = 2.3$; $p = 0.04$) were marginally significantly and significantly above the boundary line when tested across blocks. Error rates for blue were significantly below the boundary line, hence contradicting any category effect ($t(14) = -5.8$; $p < 0.001$).

In the second group (Figure S10.c-d), relative response times for the five categories pink, orange, yellow, green, and purple were all significantly above the block-wise boundary lines (min. $t(14) = 3.9$, all $p < 0.01$). Relative error rates also lay significantly above the boundary lines for pink, orange and green (min. $t(14) = 5.5$, all $p < 0.001$) and marginally significantly for yellow ($t(14) = 1.9$, $p = 0.08$). In contrast, response times ($t(14) = -6.0$; $p < 0.001$) and error rates ($t(14) = -5.8$, $p < 0.001$) for blue were below the block-wise boundary lines, hence contradicting (again) a category effect.

In sum, the t-tests across blocks confirm the effects found for average response times and error rates (Figure 4 and section "Main results: Category Effects" of the main article).

Lateralization

When participants fixate the center of the screen the left side of the screen corresponds to the left, the right side to the right visual field. In our set-up, two of the four colored disks were on the left, the other two on the right side of the screen, and participants had to fixate the center in particular since the lightness of the fixation point gave feedback about the accuracy of each response. We reanalyzed our data by comparing trials in which the target was on the left to those where it was on the right side of the screen. In the case of lateralized category effects, category effects are stronger in the right than in the left visual field. In this case, response times and error rates should decrease more strongly towards the boundaries in the right than in the left visual field.

In the first two subsection, we compare category effects between the two visual fields by visual inspection. Then, we provide statistical tests in the sections "Tests across participants (Fig. S13)" and "Tests across blocks (Fig. 14, Tab. S5)". Finally, we inspect how lateralization effects develop over time in the section "Lateralization across time (Tab. S6)".

Lateralization for group 1 (Fig. S11)

Figure S11 shows response times (upper row, panels a & c) and error rates (lower row, panels b & d) of the first group separately for the left (left column, panels a & b) and right visual field (right column, panels c & d). By visual inspection, patterns of category effects occurred in the right, but barely in the left visual field for orange response times, and for pink, orange, and purple error rates. For green, response times and error rates were above the boundary line on both sides, but more on the right side.

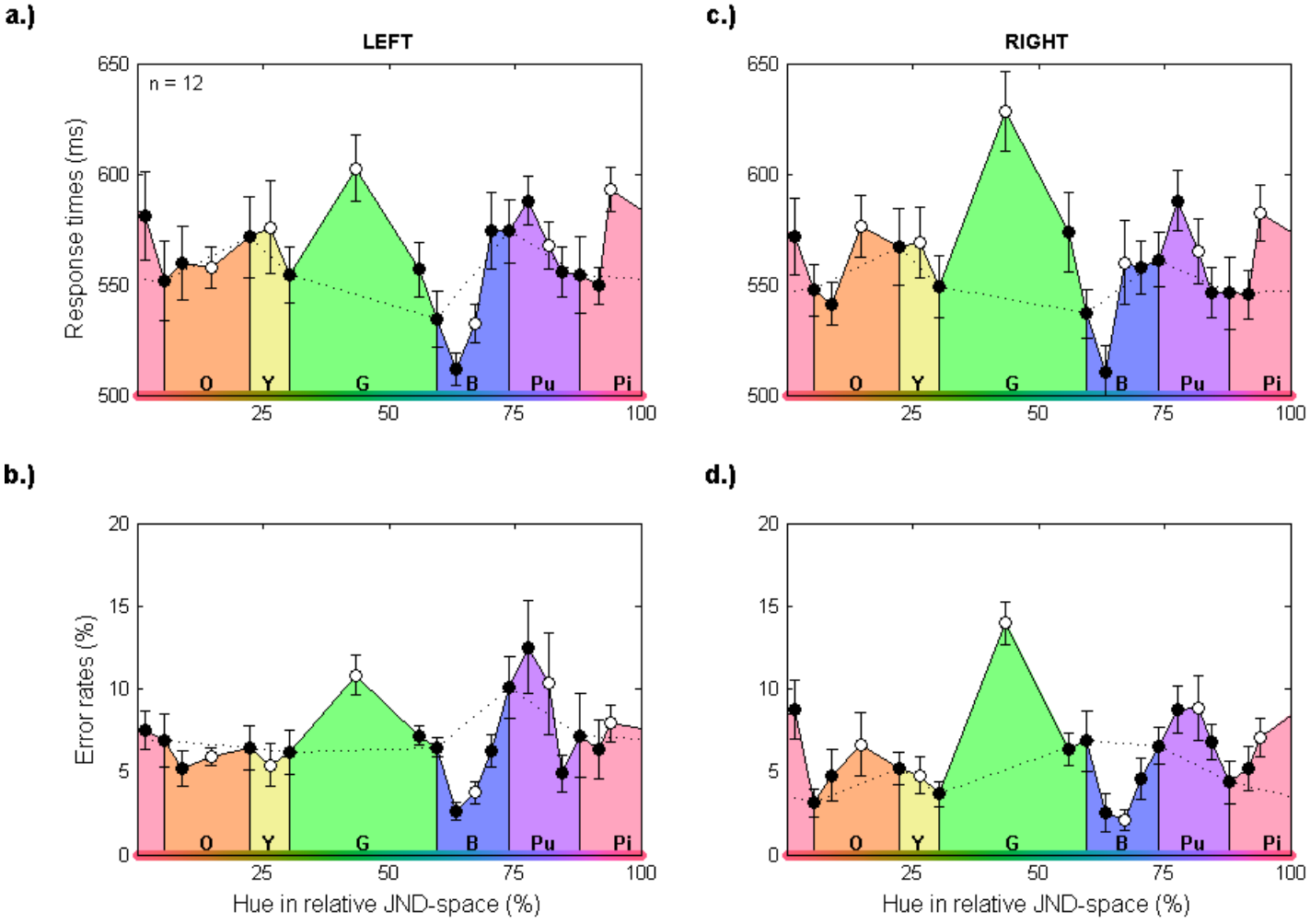

Figure S11. Lateralized performance of group 1. Format is as in Figure 4 of the main article. The only difference is that data is shown separately for trials where the target (odd one) appeared in the left (panels **a & b**) and right visual field (**c & d**). *Note that in some cases response times and error rates were above the boundary line on the right, but barely on the left side, namely for orange response times, and pink, orange, and purple error rates. For green, response times and error rates were above the boundary line on both sides, but stronger so on the right side.*

Lateralization for group 2 (Fig. S12)

Figure S12 shows the lateralized data for the second group. In general, the decrease of response times and error rates towards the boundaries of pink, orange, yellow, green, and, in tendency, for purple, seem to occur in both visual fields.

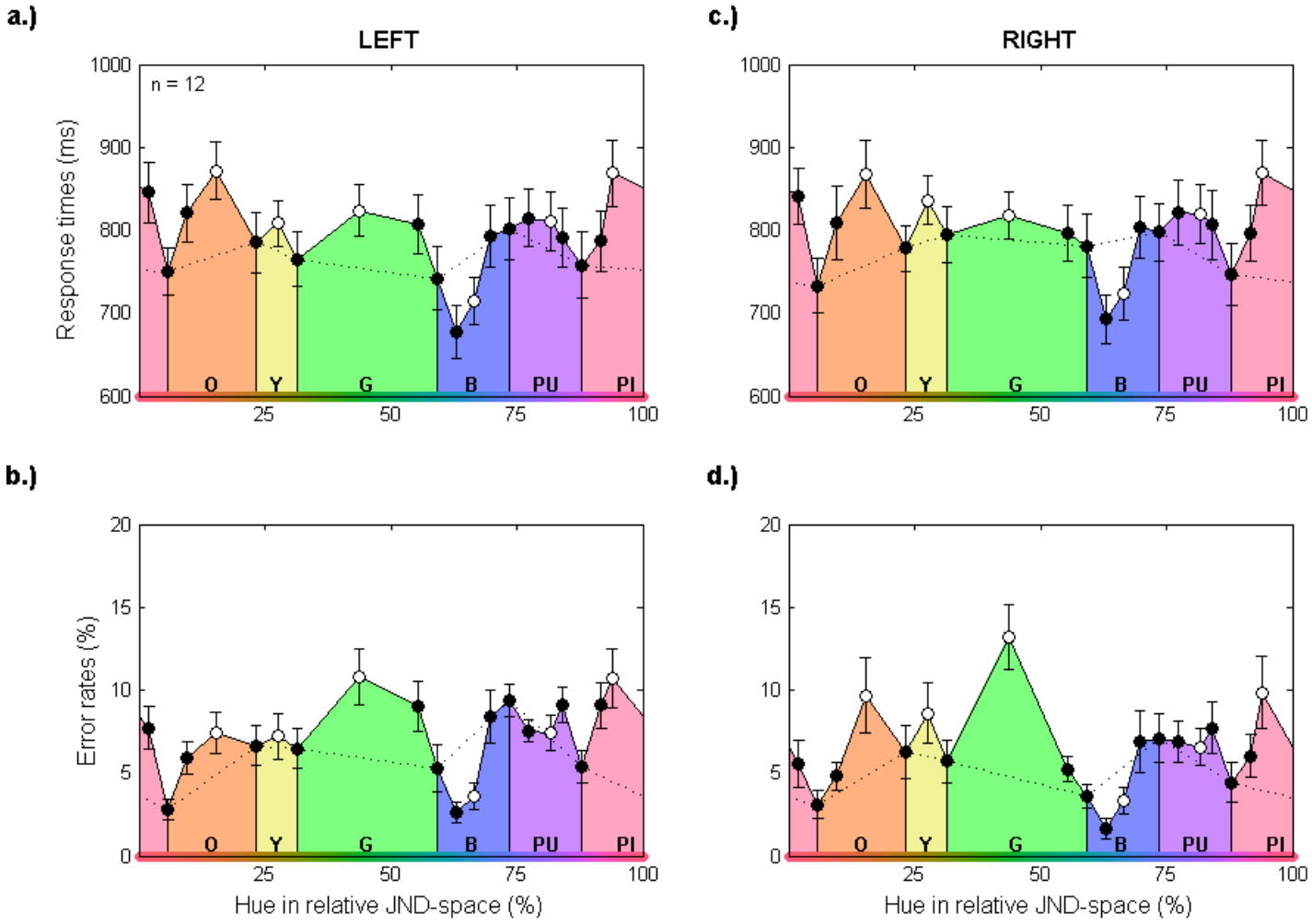

Figure S12. Lateralized performance of group 2. Format is as in Figure S11. *Note that in general the profiles looked very similar for the left and right side.*

Tests across participants (Fig. S13)

To test whether category effects were stronger in the right than in the left visual field, we calculated *laterality contrasts*. Laterality contrasts are the differences between the response times (error rates) on the right and on the left side. Figure S13 illustrates the laterality contrasts of response times (upper row, panels a & c) and error rates (lower row, panels b & d) for the first (left column, panels a & b) and second group (right column, panels c & d). If there were stronger category effects in the right visual field, the contrasts should decrease towards the boundaries.

As for the main tests for category effects, we used paired t-tests to test whether the laterality contrasts for center pairs lay significantly above the boundary lines of the laterality contrasts (dotted lines in Figure S13). Detailed results of these t-tests are provided in the upper part of Table S5 (sections a-b).

In group 1 (Figure S13.a-b), the center pairs of orange, green, blue, and purple yielded average response time and error rate contrasts above the boundary line. For blue this was true for response times, and for yellow for error rates, only. However, none of these differences reached significance across observers (min. $p = 0.15$). Only green error rate contrasts were marginally significant above the boundary line ($t(8) = 2.0$, $p = 0.08$).

In group 2 (Figure S13.c-d), the center pairs of pink, orange, yellow, and purple resulted in response time contrasts above the boundary line. For error rates this was the case for orange, yellow, green, blue, and purple. However, none of these differences from the boundary line were significant (min. $p = 0.15$). The response time contrasts for green yielded a marginally significant pattern ($t(11) = -2.1$; $p = 0.07$) that contradicted the lateralized category effect, in that category effects tended to be stronger on the left side.

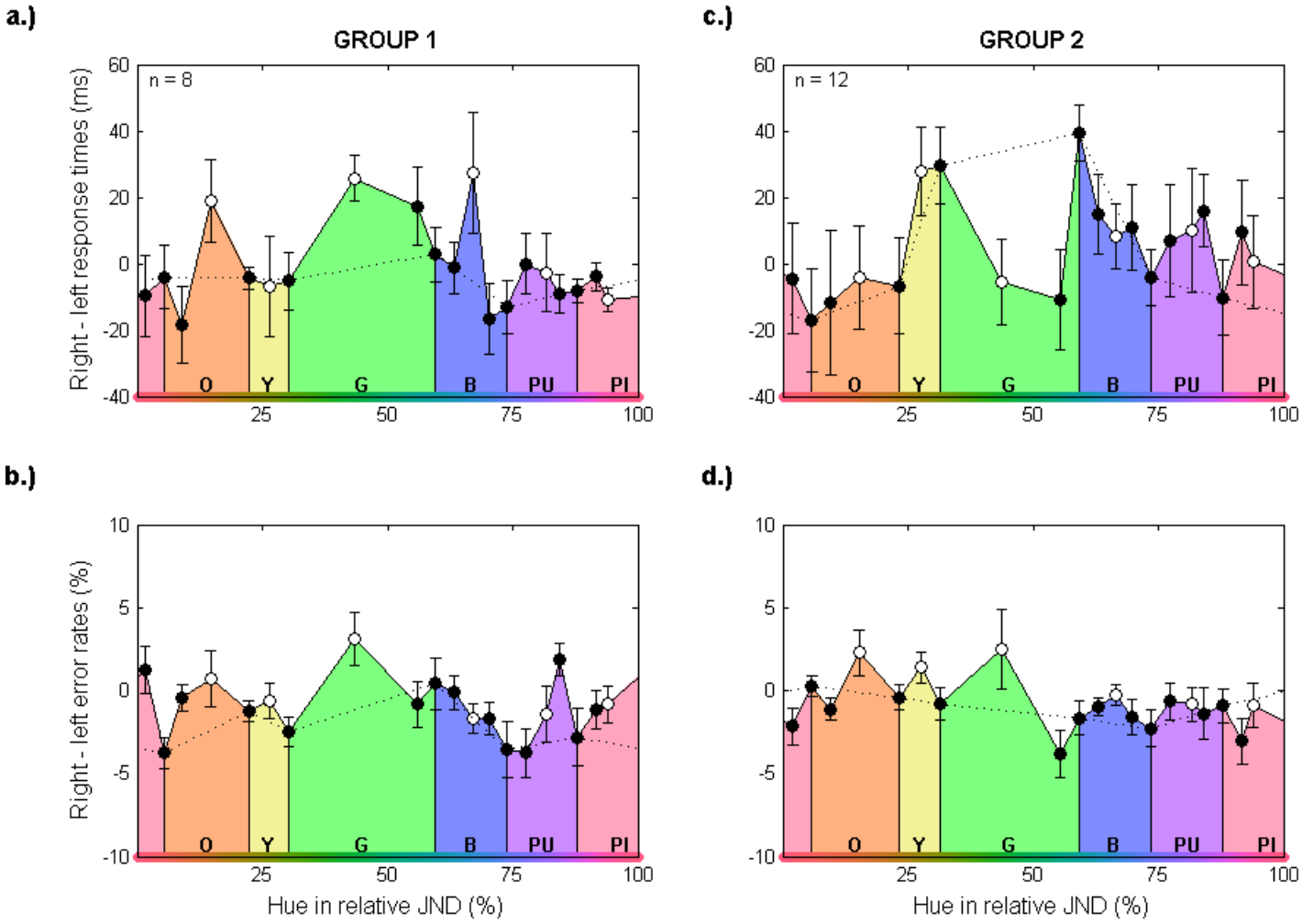

Figure S13. Laterality contrasts between right and left. The y-axis represents response time (panels **a & c**) and error rate (**b & d**) laterality contrasts for the first (**a & b**) and the second group (**c & d**). These contrasts correspond to the difference between the relative response times (error rates) on the left and those on the right side. Apart from that, format as in Figure 4 of the main article. In the case of a lateralized category effect, boundary pairs should yield lowest, center pairs highest laterality contrasts, indicating stronger category effects on the right than on the left side. *Note the non-significant tendencies towards this pattern for orange, green, and purple in the first group, and for orange, yellow, and purple in the second group, as well as the opposite pattern for green response times in the second group (panel c).*

Tests across blocks (Fig. 14, Tab. S5)

The lack of significant results in the t-test across participants might have been due to low statistical power because of the limited number of participants. To further explore the non-significant tendencies towards a lateralization effect we tested for lateralization effects across the 15 blocks of the measurements. For this purpose, we determined laterality contrasts of relative response times and error rates for each block. Like above, the laterality contrasts are calculated as the difference between the relative response times (error rates) for the center pair on the right and on the left side. The resulting laterality contrasts are shown in Figure S14. We tested whether these laterality contrasts were above zero through paired, two-tailed t-tests across blocks. Results are summarized in the lower part of Table S5 (sections c-d).

In the first group (Figure S14.a-b, Table S5.c), orange showed a significant and marginally significant lateralisation pattern for error rates ($t(14) = 3.1$, $p = 0.008$) and response times ($t(14) = 2.0$, $p = 0.07$), respectively. As in the tests across participants, this result indicates that relative response times and error rates were higher above the boundary line in the right than in the left visual field. Blue yielded a contrary lateralisation effect on response times ($t(14) = 2.4$, $p = 0.03$), which implied that response times were less below the boundary line on the right than on the left side.

In the second group (Figure S14.d, Table S5.d), only the lateralisation effect for green error rates was significant ($t(14) = 2.4$, $p = 0.03$), the one for orange error rates was only marginally significant ($t(14) = 1.8$, $p = 0.09$). There was no other significant lateralisation effect across blocks, neither for response times (min. $p = 0.19$), nor error rates ($p = 0.22$).

In sum, there were some tendencies towards a (right-) lateralized category effect in some categories. However, there was no consistent lateralization of category effects in either group.

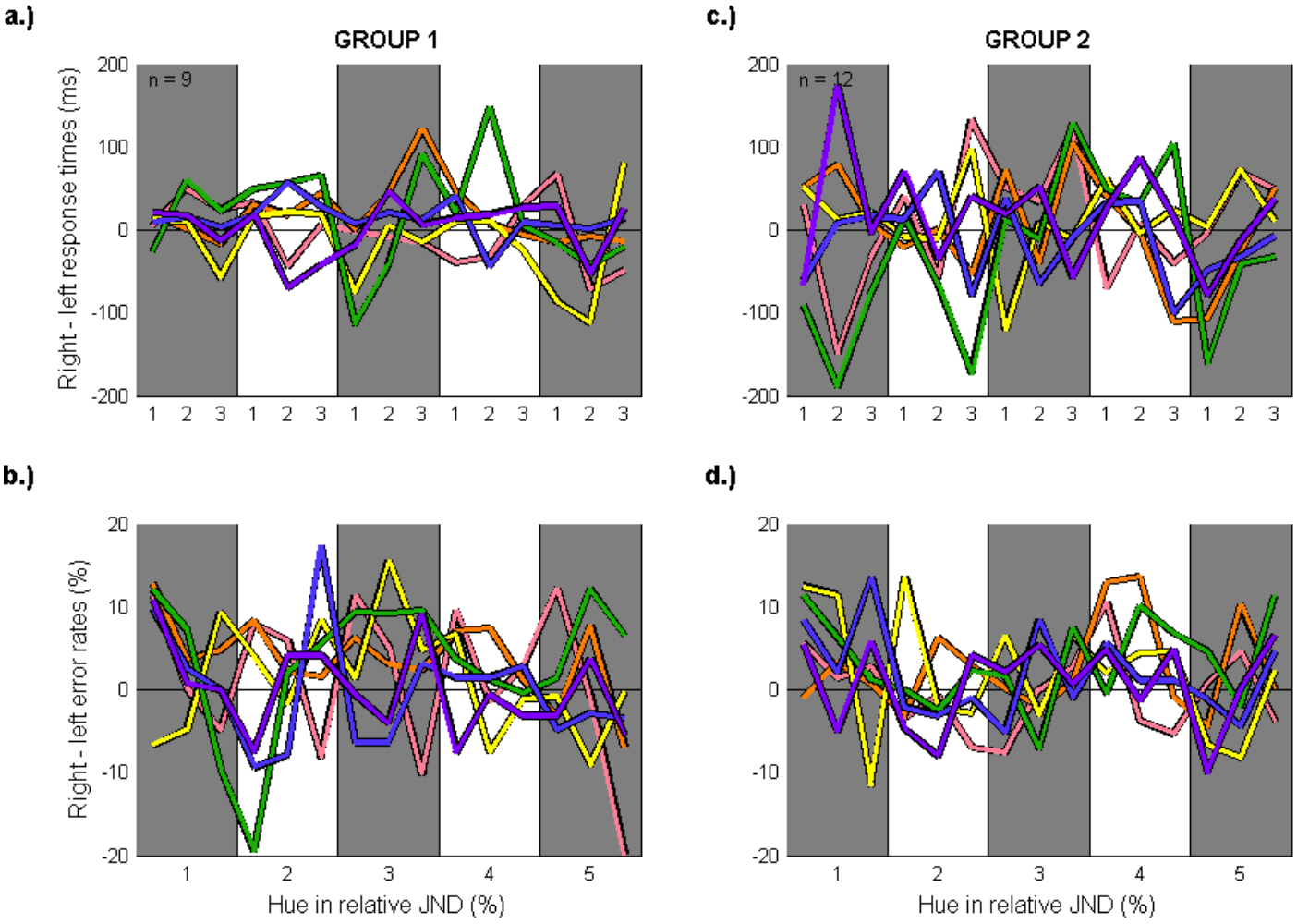

Figure S14. Lateralization effects over time. The y-axis represents the laterality contrasts for the relative response times (panels **a & c**) and error rates (**b & d**) of the first (**a & b**) and second group (**c & d**). These contrasts are calculated as the difference between the relative response times (error rates) for the center pairs when the target was left and when it was right to the fixation point. Apart from that, format as in Figure S10. *Note that there is no systematic pattern across blocks in any of the groups, with the exception of the negative correlation for the orange error rates in the first group.*

		Response times (in ms)			Error rates (in %)				Response times (in ms)			Error rates (in %)		
Category	*df*	*M*	*t*	*p*	*M*	*t*	*p*	*df*	*M*	*t*	*p*	*M*	*t*	*p*
a.) Group 1 (trained): t-test across participants								**c.) Group 2: t-test across participants**						
Pink	8	-0.6	-0.1	0.95	1.9	0.8	0.46	11	13.1	0.9	0.37	-0.4	-0.2	0.83
Orange	8	24.6	1.6	0.15	4.0	1.6	0.15	11	7.3	0.4	0.73	2.4	1.4	0.18
Yellow	7	-0.9	0	0.97	1.1	0.7	0.54	11	15.6	0.8	0.42	2.0	1.6	0.15
Green	8	24.4	1.6	0.14	3.6	2.0	0.08	11	-39.3	-2.1	°	3.6	1.3	0.23
Blue	8	29.3	1.1	0.32	-0.5	-0.3	0.79	11	-9.2	-0.6	0.58	1.7	1.3	0.23
Purple	8	1.8	0.1	0.92	1.6	0.6	0.59	11	17.7	0.8	0.42	0.7	0.4	0.67
b.) Group 1 (trained): t-test across blocks								**d.) Group 2: t-test across blocks**						
Pink	14	-6.3	-0.3	0.78	1.4	0.6	0.57	14	12.0	0.6	0.54	-0.2	-0.2	0.87
Orange	14	18.2	2.0	°	3.9	3.1	**	14	3.0	0.2	0.86	2.6	1.8	°
Yellow	14	-12.3	-1.0	0.36	1.2	0.7	0.52	14	14.1	1.1	0.28	1.9	1.0	0.34
Green	14	17.2	1.0	0.32	3.4	1.5	0.15	14	-33.8	-1.4	0.19	3.4	2.4	*
Blue	14	13.4	2.4	*	-0.2	-0.1	0.92	14	-12.6	-1.0	0.34	1.8	1.3	0.22
Purple	14	1.8	0.2	0.84	0.1	0	0.98	14	17.7	1.0	0.31	0.7	0.5	0.63

Table S5. T-test for lateralized category effects. The paired, two-tailed t-tests compared differences between center and boundary pairs on the right and left side. The left part of the table (**a & b**) report results for the first, the right part (c and d) for the second group; the first row (a and c) gives statistics for tests across participants, the second row (b and d) across the 15 blocks.

Lateralization across time (Tab. S6)

If lateralization effects depend on training, lateralization effects might be covered by lumping all blocks together. For this reason, we inspected whether potential lateralization effects depend on time and training. For this purpose, we calculated correlations between the block number (1 to 15) and the average lateralization effects per block. Detailed results are provided in Table S6.

In group 1 (Table S6.a), only the lateralization of the relative error rates for orange correlated negatively with blocks ($r(15) = -0.52$, $p = 0.045$). There was no other significant correlation, neither for response times, nor error rates. In group 2 (Table S6.b), correlations between lateralization and blocks were not significant for any of the categories, neither for response times (min. $p = 0.20$) nor for error rates (min. $p = 0.19$). Hence, there was no consistent modulation of lateralization effects across time, which further supports the conclusion that there were no consistent lateralization effects.

		a.) Group 1 (experienced)				**b.) Group 2 (non-experienced)**			
		Response times (in ms)		Error rates (in %)		Response times (in ms)		Error rates (in %)	
Category	*n*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Pink	15	-0.40	0.14	-0.29	0.30	0.24	0.38	-0.09	0.76
Orange	15	-0.14	0.62	-0.52	*	-0.34	0.22	0.21	0.45
Yellow	15	-0.17	0.53	-0.16	0.57	0.06	0.83	-0.36	0.19
Green	15	-0.14	0.62	0.22	0.43	0.35	0.20	0.10	0.74
Blue	15	-0.26	0.36	-0.21	0.45	-0.19	0.50	-0.29	0.30
Purple	15	0.12	0.67	-0.35	0.20	-0.18	0.52	0.06	0.84

Table S6. Correlations between lateralization and time. Part **a** (left) shows results for the first, part **b** (right) those for the second group. "r" reports the correlation coefficient, and "p" the corresponding two-tailed t-statistic.

VALIDATION OF CATEGORIES

Naming test of main experiment

The naming test of the main experiment allowed for assessing differences between the actual categories for the 32 colors of the speeded discrimination task, and the assumed categories measured for the 120 colors in the preliminary naming test. Figure S15 and Figure S16 provide trial-by-trial results for each participant. They correspond to the aggregated data in Figure 6 of the main article.

Differences between groups (Fig. S15)

Rationale: If the discrepancies between assumed and actual categories were stronger in the first than in the second group, this would explain why category effects were more pronounced in the second than in the first group.

Results: In both groups the category boundaries of the naming test of the main experiment did not completely agree with the boundaries of the preliminary measurements. In group 1 (Figure 6.a of the main article) original boundaries (dotted lines) correspond to the individual boundaries of each observer. In this group, red, orange, yellow, and green colors deviate particularly from the assumed boundaries.

In group 2 (Figure 6.b of the main article), the boundaries of the original categories (dotted lines) correspond to the aggregated categories of group 1. Differences between the two kinds of measurements appeared in all observers and concerned all categories.

Finally, Figure S15 and Figure S16 also show that category membership of colors varied across sessions. This is particularly true for colors around the boundaries.

Discussion: Together, these results suggest that categories might have differed between the different stimulus sets (32 vs. 120 colors). Hence, there might have been discrepancies between the assumed and the actual categories for the equally discriminable colors. These discrepancies could potentially explain why there were no categorical facilitation effects in the first group.

However, the fact that there were also such discrepancies for the second group is at odds with the presence of category effects in the second group. Hence, the discrepancies between assumed and actual categories cannot explain the differences in category effects between the two groups. Moreover, in both groups some of the observed differences between measurements might just be due to intra-individual variation across sessions (Figure S15 & S16).

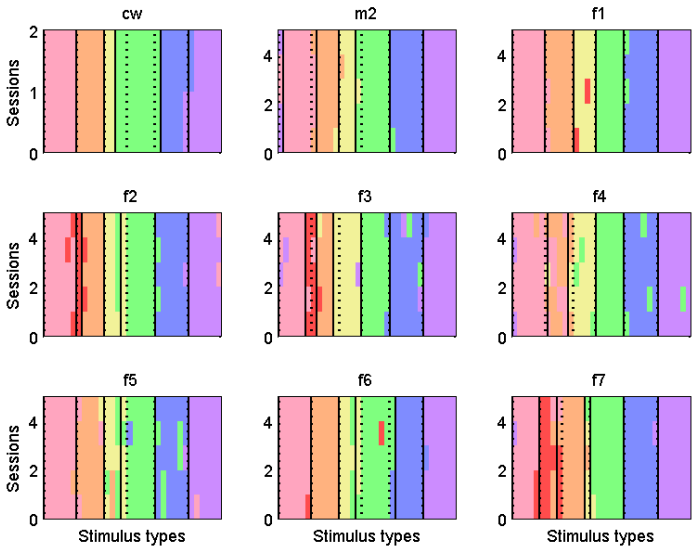

Figure S15. Individual results of group1 in the naming test of the main experiment. In each panel, the x-axis corresponds to the 32 colors of each stimulus set, the y-axis to the sessions. Boundaries based on the preliminary measurements by Witzel and Gegenfurtner (2013) are shown as dotted black lines. Boundaries updated through the present data are shown as solid black lines. Apart from that format as in Figure 6 of the main article. *Note the deviations of the re-measured category boundaries from the assumed boundaries.*

Blue-green boundary (Fig. S16)

Rationale: Systematic discrepancies at the green-blue boundary would explain the contradictory pattern at this boundary. What contradicted the category effect was the observation that the transitional blue-green pair rather than the green-blue boundary pair yielded maximal performance in both groups (cf. Figure 4 of main article). However, the local minimum at the transitional blue-green pair would be in line with a category effect if the actual green-blue boundary would coincide with this pair rather than the assumed green-blue boundary pair. To explain the increase in performance at the blue-green transitional pair in the speeded discrimination task, the boundary in the present naming test should be consistently shifted towards this blue-green transitional pair.

Results: However, the green-blue boundary did not vary a lot between the preliminary measurements and those in the main experiment (cf. Figure 6). Apart from only two such shifts in each group, there were also two shifts towards the opposite direction in the second group that is towards the green-blue transitional pair (cf. Figure 6.b). Consequently, discrepancies between actual and assumed category

boundaries cannot explain the absence of category effects in the blue category.

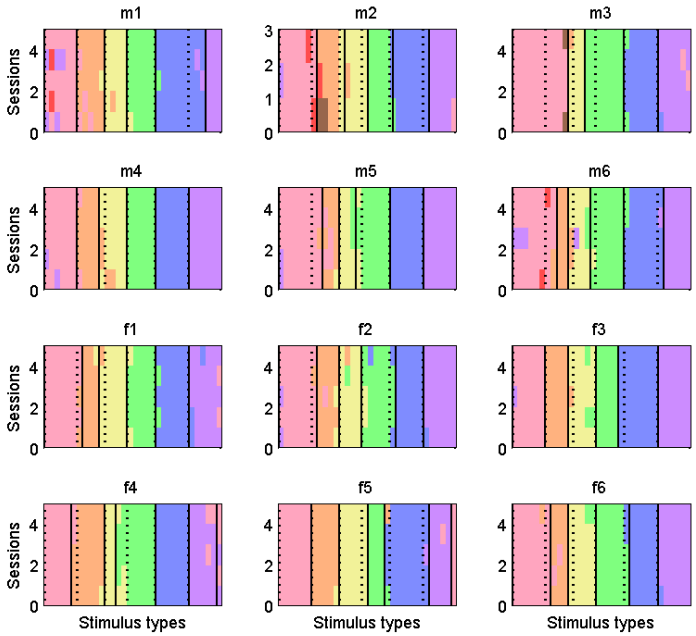

Figure S16. Individual results of group 2 in the naming test of the main experiment. Format as in Figure S15. *Note that differences between assumed (dotted black lines) and measured (solid black lines) may be due to differences between the two groups and to differences in color samples between preliminary measurements and measurements during the main experiment.*

Post-hoc naming test (Fig. S17)

Rationale: For the second group, discrepancies between the original and the re-measured categories were not necessarily due to differences in stimulus sets. They may also result from individual differences in color naming because the aggregated color categories of the first group were used for the creation of equally discriminable stimuli of the second group. The post-hoc color naming with the 6 observers of the second group was done with the same stimulus set of 120 colors as the preliminary naming of Witzel and Gegenfurtner (2013). Hence, it allowed for assessing differences between the aggregated categories of the first group and the individual categories of the second group.

Results: Figure S17 shows the results of the post-hoc color naming test across the whole hue circle. The colored areas represent the mode color names across the 6 sessions for each individual and each of the 120 colors along the hue circle. Dashed lines show the aggregated category boundaries measured preliminarily by Witzel & Gegenfurtner (2013) for the first group of participants. Black and colored disks represent the equally discriminable stimuli of the main experiment. Colored disks highlight stimuli for which category membership differs between the aggregated categories of the first group and the individual categories of the second group. Their colors indicate the membership to the aggregated categories of the first group. The comparison between those discs and the measured categories (colored areas and thick black lines) allows for evaluating whether differences between the two measurements are important enough to affect the category membership of the equally discriminable colors. Indeed, there were some differences between the preliminary and the post-hoc categories. This was the case for the yellow colors of m4, the pink-orange boundary of m5, and the purple-pink boundary of f2.

Discussion: These results indicate that the individual categories of the second group differed from the aggregated categories of the first group. However, the second group with the aggregated categories yielded category effects, not the first group with the individual categories. Hence, the present results support the idea that aggregated categories (as used for producing the equally discriminable stimulus pairs for the second group), are more relevant for categorical facilitation than individual categories (as used for producing the stimulus pairs for the first group).

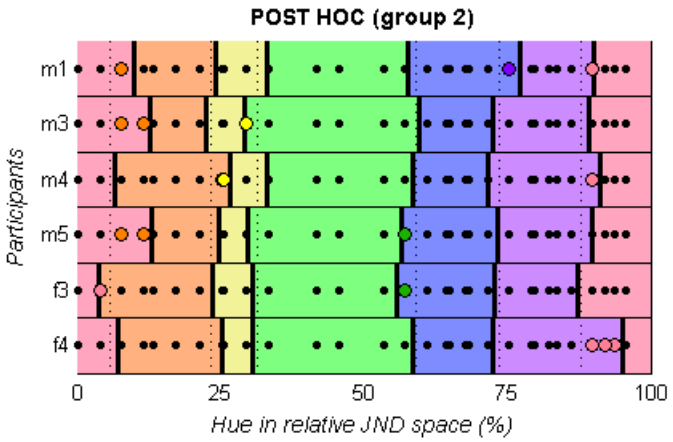

Figure S17. Post-hoc categories of group 2. The graphic illustrates the result for the post-hoc measurements of all colors along the hue circle for 6 observers of the second group. Format as in Figure 6 of the main article, except for the x-axis, which represents the hue circle in relative JND-space as in Figure S2.b (and Figure 4). *Note that the post-hoc categories (colored areas and thick solid lines) slightly deviate from the assumed categories (disks and dotted lines).*

Re-categorization of stimulus pairs (Fig. S18)

Rationale: The re-measurements of color categories indicated that the actual categories for the set of equally discriminable colors might differ from those assumed for the production of the equally discriminable color pairs. The question arises whether the differences between preliminary and actual categories were strong enough to affect our main results on categorical facilitation. In particular, the observed differences in categorization suggest that there were differences between individual and aggregated categories. These differences might be important for category effects, and could explain the differences in category effects

between the first and second group. Moreover, the difference between assumed and actual categories might account for the contrary effects in the blue category.

Method: We tested these ideas by re-categorizing the equally discriminable stimulus pairs. Re-categorization means that the equally discriminable color pairs are reassigned to the different types of color pairs (boundary, transitional, and center pairs) according to a set of categories other than those originally assumed for the production of these stimulus pairs.

However, the original stimulus pairs were matched to the assumed, preliminary categories. For this reason, the re-categorization according to an alternative set of categories that differed from those assumed categories could not provide pairs of each type (i.e. center, transitional, and boundary pairs). For this reason, no distinction between center and transitional pairs was possible, and stimulus pairs were re-categorized as boundary pairs and within-pairs. The latter referred to any pair within a category. Boundary pairs were determined as pairs with a boundary anywhere between the colors. The colors of the within-pairs lay both within the boundaries of the same category. Still, this re-categorisation involved a reduction in the number of participants for each comparison because for some participants the alternative boundaries would require new color pairs, for which no speeded discrimination data was available.

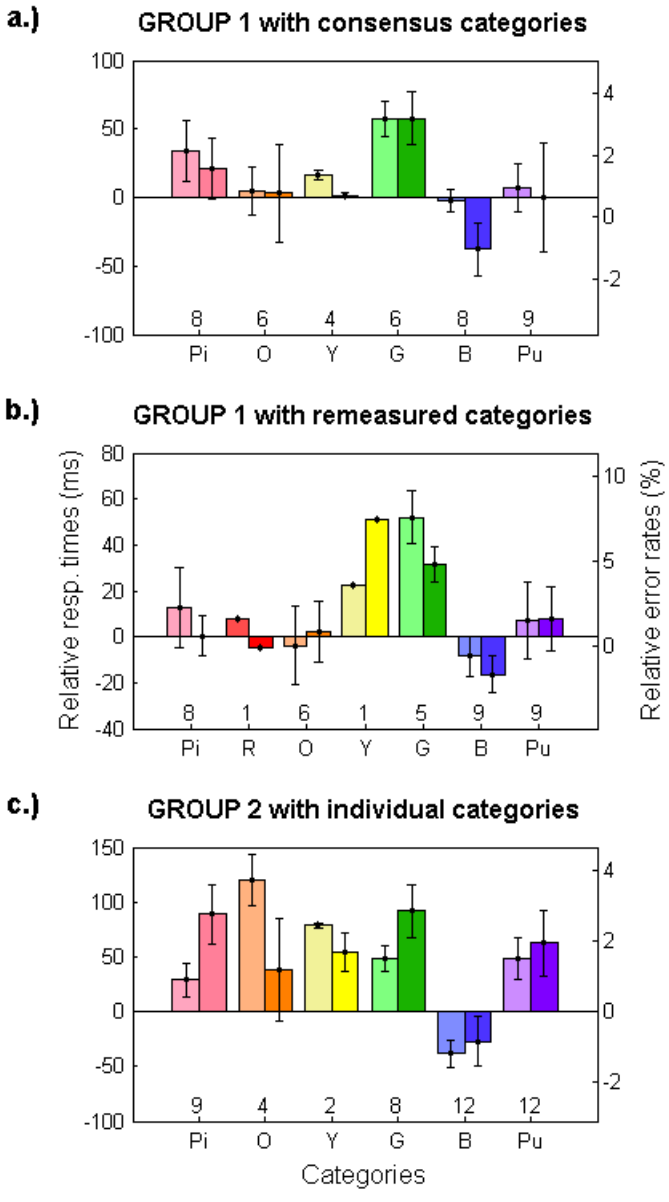

Figure S18. Category effects for re-categorized color pairs. Graphics show relative response times (left y-axis) and error rates (right y-axis) for re-categorized stimulus pairs. In panel **a**, individual stimulus pairs of the single participants of the first group were re-categorized according to the consensus categories, i.e. the average category boundaries of this group. In panel **b**, the stimulus pairs of the first group were re-categorized by the category boundaries measured through the naming test in the main experiment. In panel **c**, the stimulus pairs of the second group were re-categorized according to the individual categories of this group, as measured in the naming test of the main experiment. The colors of the bars refer to the different categories, the initials of which are listed along the x-axes. Light, desaturated bars refer to response times, dark saturated ones to error rates. The number of participants with a set of re-categorized stimuli is indicated above the x-axis. *There were no new categorical patterns for the consensus categories and the re-measured categories in the first group (panel a and b), and not less categorical patterns for the individual categories in the second group (panel c). Moreover, the blue category showed the inverse pattern independently of the set of categories.*

←

Results: Figure S18 illustrates the results of the re-categorizations. In all panels of Figure S18, The bars refer to the average distances from the boundary line of the response times (light, desaturated bars, left y-axis) and error rates (dark, saturated bars, right y-axis) within the categories. Positive bars indicate that response times and error rates, respectively, were higher within than across categories, as predicted by categorical facilitation. For several participants there were no boundary pairs for some categories. The number of remaining participants for each category are shown at the bottom of the graphics.

The lack of category effects in the first group might be due to the use of individual instead of aggregated categories. Consequently, more or stronger category effects should appear when re-categorizing the stimuli according to their consensus categories, i.e. those categories that were used for the production of equally discriminable color pairs for the second group. At the same time, if individual categories are detrimental for category effects, category effects

in the second group should disappear or at least decrease when their data is re-categorized by their individual categories. Finally, we inspected whether the blue category yielded category effects when re-categorizing the stimulus pairs by other measurements of categories.

First, we examined whether aggregated categories produce more categorical patterns in the first group. For this purpose, we re-categorized the individual stimulus sets of the first group according to the aggregated categories of this group, i.e. the same boundaries that were used for the creation of the second group's equally discriminable stimuli (Figure S18.a). Second, we inspected whether the re-measured categories in the naming task of the main experiment produced stronger categorical patterns in the first group. For this purpose, we re-categorized the stimulus sets of the first group according to the new categories obtained in the naming test of the main experiment for this first group (Figure S18.b). Finally, we examined whether the second group would yield less category effects when using individual categories. Hence, we re-categorized their stimuli according to their individual categories from the naming test of the main experiment (Figure S18.c).

In general, no new patterns appeared in any of the two groups after the re-categorizations. In the first group (Figure S18.a-b), patterns of categorical facilitation appeared for green response times and error rates. Apart from that, there were some non-significant tendencies for pink, and a significant inverse pattern for blue error rates. In the second group (Figure S18.c), the bars of the same 5 categories that showed categorical facilitation with the original categories (pink, orange, yellow, green, and purple) were positive for response times and error rates, while those for blue were negative.

Discussion: If category effects were stronger for aggregated than for individual categories, re-categorization should enhance category effects in the first and attenuate them in the second group. However, neither has been the case in the present study. On the one hand, the application of aggregated categories did not reveal additional or at least stronger category effects in the first group (Figure S18.a). On the other hand, the categorical facilitation effects for all 5 categories still appeared in the response times and error rates when the data of the second group was re-categorised by the individual categories of the naming test of the main experiment (Figure S18.c). These results show that the difference between individual and aggregated categories cannot explain the different category effects in the first and second group.

Finally, no categorical patterns appeared for blue when re-categorizing the stimuli. This result shows that the contradictory pattern in the blue category cannot be due to systematic discrepancies between assumed and actual categories.

JNDs AND SPEEDED DISCRIMINATION

JNDs

JNDs and speeded discrimination (Tab. S7)

1. Measure	*2. Measure*	*N*	*r*	*p*	R^2
a.) Correlation between JNDs (Fig. 3.a)					
Pre-JNDs G1	Post JNDs G2	20	0.91	***	83%
b.) Pre-JNDs and speeded discrimination (Fig. S20)					
Pre-JNDs G1	Speed RTs of G1	20	0.13	0.58	2%
Pre-JNDs G1	Speed ERs of G1	20	0.16	0.49	3%
Pre-JNDs G1	Speed RTs of G2	20	-0.15	0.52	2%
Pre-JNDs G1	Speed ERs of G2	20	0.24	0.31	6%
c.) Post-JNDs and speeded discrimination (Fig. 20)					
Post-JNDs G2	Speed RTs of G1	20	-0.31	0.19	10%
Post-JNDs G2	Speed ERs of G1	20	-0.17	0.47	3%
Post-JNDs G2	Speed RTs of G2	20	-0.34	0.14	12%
Post-JNDs G2	Speed ERs of G2	20	-0.12	0.62	1%
d.) JND differences (Fig. S19)					
JND diff (G1-G2)	Speed RTs of G2	20	-0.68	***	47%
JND diff (G1-G2)	Speed ERs of G2	20	-0.72	***	52%

Table S7. Pre- and post-JNDs. Correlations between JNDs and performance in the speeded discrimination task (G1 = Group 1; G2 = Group 2). Part **a** reports the correlation between the preliminary (Pre-JNDs) and post-hoc JNDs (Post-JNDs); part **b** the correlations between preliminary JNDs and performance in the speeded discrimination task (Speed RT = response times, speed ER = error rates in speeded discrimination task); part **c** correlations between the post-hoc JNDs and performance in the speeded discrimination task; and part **d** correlations between the differences between preliminary and post-hoc JNDs on the one hand, and performance of the second group in the speeded discrimination task on the other.

Table 7.a reports the correlations between preliminary and post-hoc JNDs (black and red line in Figure 3.a). As reported in the main article (section "JNDs"), the strong correlation shows the similarity between the two JND measurements.

Figure S20 allows for comparing JNDs (thick black line) and response times in the speeded discrimination task (thin black line). Correlations across the 20 equally discriminable color pairs were calculated between JNDs and the response times and error rates in the speeded discrimination task. For the first group (Figure S20.a), JNDs for the 20 centroids of the 20 color pairs were interpolated based on the original measurements for the 72 test colors. Results are summarized in Table S7.b-c. Preliminary and post-hoc JNDs were not correlated to the performance in the speeded discrimination task for any of the two groups (among all 8 correlations: max. R^2 = 12%, min. $p = 0.14$). As reported in the main article (section "JNDs"), these results show that the patterns of JNDs across hues differed from the patterns of response times and error rates in the speeded discrimination task.

Differences between preliminary and post-hoc JNDs (Fig. S19)

We assessed the potential impact of differences between preliminary and post-hoc JNDs on the control of discriminability in the speeded discrimination task. The color pairs in the speeded discrimination task differed by 2 JNDs according to the preliminary JNDs of the first group. We recalculated the differences of two JNDs based on the aggregated post-hoc JNDs of the second group.

Results are shown in Figure S19.a. The x-axis corresponds to hue angle in DKL-space (for comparison with Figure 3.a), and the y-axis corresponds to equally discriminable differences between the two colors of a stimulus pair, measured in numbers of JNDs. The horizontal black line shows the equally discriminable differences between the colors in each pair when determined by the preliminary JNDs (i.e. those shown by the black line in Figure 3.a). Due to the production of the color pairs, the difference shown by this black line is 2 JNDs for all stimulus pairs. The solid black curve above the colored area, refers to the equally discriminable differences determined by the post-hoc JNDs (i.e. those shown by the dashed red line in Figure 3.a). Each data point on this curve corresponds to the center of one of the 20 color pairs. If this curve is higher than 2 JNDs, this implies that the difference between the colors of the respective color pair would be larger than 2 JNDs when re-evaluated with the post-hoc JNDs, and vice versa.

In general, the curve for the post-hoc JNDs varied around the 2-JND line of the preliminary JNDs, indicating that the two measurements yielded more or less the same absolute differences. However, at the pink-orange, yellow-green, and purple-pink boundary, the curve for the post-hoc JNDs is higher, and for the yellow, green, and pink center it decreases. Moreover, around the blue-green transitional pair the post-hoc measurements were particularly high. This implies that there were some tendencies of this curve to be high where response times (green curve) and error rates (red curve) were low in the speeded discrimination task and vice versa.

These observations suggest that the stimulus pairs that yielded comparatively low performance in the speeded discrimination task (Figure 4) also yielded lower sensitivity (higher JNDs) in the post-hoc JND measurements as compared to the preliminary JND measurements.

In particular, the differences between the JND measurements were in line with patterns of performance that were attributed to category effects. Colors at the category centers (pink, yellow, green, maybe purple) resulted in comparatively low performance in the speeded discrimination task. These colors were also less discriminable in the post-hoc JND measurements than predicted by the preliminary measurements. This is shown by the fact that distances of 2 JNDs were particularly low for those colors when measured by the post-hoc instead of the preliminary JNDs (Figure S19.a).

In addition, they were also coincident with the patterns in the blue category that contradicted category effects. In particular, the blue-green transitional and the blue center pair yielded particularly high performance in the speeded discrimination task. According to the post-hoc measurements, they were also more discriminable than predicted by the preliminary JNDs, as illustrated by the fact that distances of 2 JNDs are larger when measured by the post-hoc instead of the preliminary JNDs (Figure S19.a).

We assessed the strength of the relationship of the differences between the two JND measurements to the discrimination performance in the speeded discrimination task. For this purpose, we calculated the correlations between those JND differences (black solid curve in Figure S19.a) on the one hand, and the response times (green dashed curve in Figure S19.a) and error rates (red dashed curve in Figure S19.a) in the speeded discrimination task. The correlations are illustrated by Figure S19.b and statistics are summarized in Table S7.d.

Indeed, the differences between the two kinds of JNDs were negatively correlated with the response times ($r(20) = 0.68$, $p < 0.001$) and error rates ($r(20) = 0.72$, $p < 0.001$) of the second group. These correlations even persist after excluding the 3 blue stimuli that contradicted a categorical pattern of response times and error rates ($r(17) = -0.57$, $p = 0.001$; $r(17) = -0.63$, $p < 0.001$). These correlations suggest a relationship between the JND measurements and speeded discrimination.

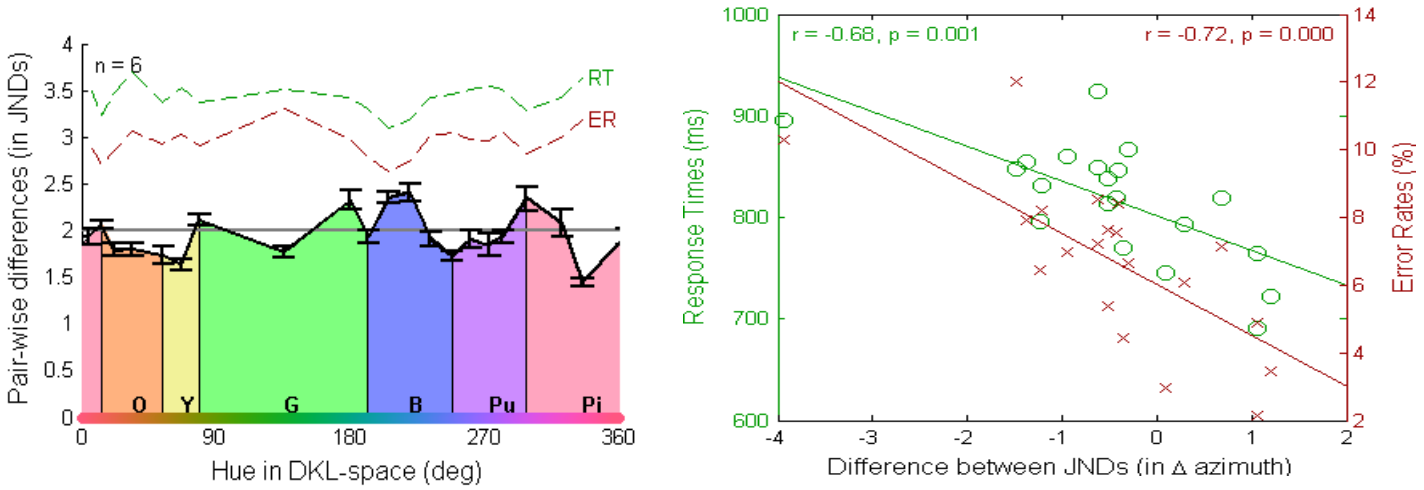

Figure S19. Differences between preliminary and posthoc JNDs. Panel **a** illustrates equally discriminable stimuli when determined by the post-hoc JNDs. The main black curve corresponds to the number of post-hoc JNDs of the second group that fit into the color differences defined by 2 JNDs of the first group. The horizontal grey line corresponds to 2 JNDs according to the preliminary JND measurements with the first group. The dashed lines show the profile of the response times ("RT", green) and error rates ("ER", red) of the second group in the speeded discrimination task (same as in Figure 4.c-d). Apart from that format as in Figure 4 of the main article. Panel **b** illustrates the correlation between the differences of the two kinds of JNDs (preliminary and post-hoc) and the response times (green disks) and error rates (red disks) of the second group in the speeded discrimination task. The lines correspond to the respective regression lines. *Post-hoc measurements (black curve in panel a) yielded slightly different estimates of discriminable differences than predicted by preliminary JNDs (horizontal black line in panel a). These differences correlated with the performance of the second group in the speeded discrimination task (panel b).*

Response times in JND measurements

Supra-threshold response times (Fig. S20)

To test for category effects during JND measurements, we determined response times for supra-threshold color differences in the JND measurements. Only response times for trials in which the difference between test and comparison was larger than 1 JND were included in the analysis. Due to the adaptive staircase, there were different amounts of data for different participants, and different test-colors. For this reason, it was impossible to control the difference between test and comparison more rigorously; for example there would not have been enough data for all observers and test colors for differences above 2 JNDs. To account for the additional variability, medians were used to aggregate response times. All 10 participants from Witzel and Gegenfurtner (2013) were included in the analysis for the first group. For each test-color and each participant, we calculated median response-times across supra-threshold comparison colors. Averages across participants are shown in Figure S20.

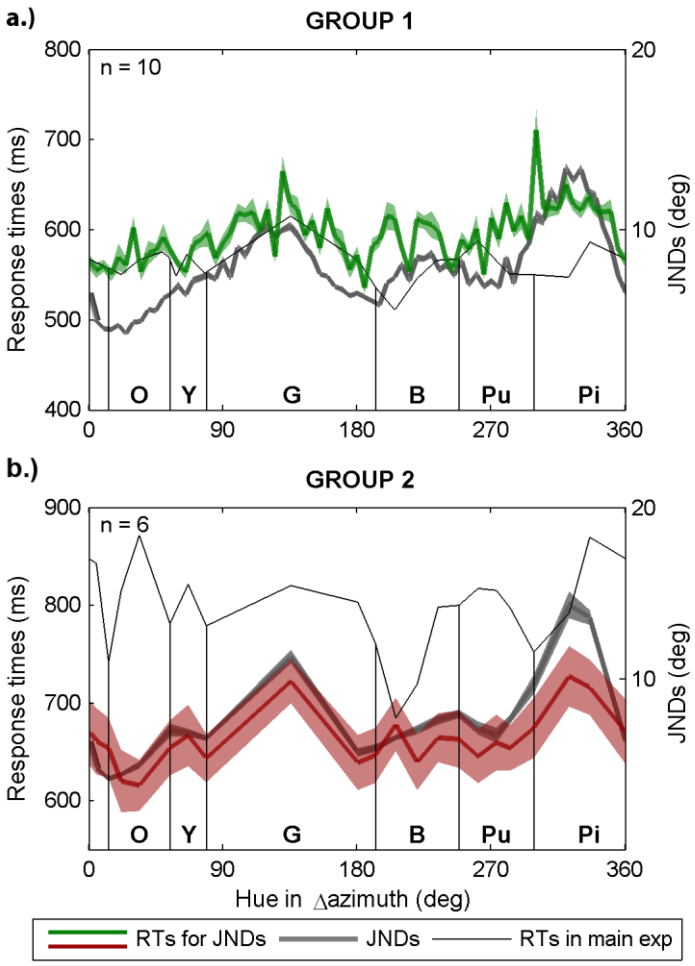

Figure S20. Supra-threshold response times in JND measurements. Panel **a** shows data for the first, panel **b** for the second group. The x-axis represents hue in azimuth degree, the left y-axis represents response times, and the right y-axis JNDs as differences in delta azimuth degree. The green (panel **a**) and red (panel **b**) curve show the average response times for supra-threshold (> 1 JND) stimuli in the JND measurements. The fat grey curves in both panels show the JNDs of the preliminary JND measurements of Witzel and Gegenfurtner (2013) and of the post-hoc JND measurements, in panel **a** and **b** respectively. They are the same as the black and the red curves in Figure 3.a of the main article. Transparent areas represent standard errors of mean across participants. The thin black curve corresponds to the response times from the speeded discriminations task of the main experiment (same as in Figure 4.a & c). Consensus categories are shown as vertical black lines, with color names indicated by their initials above the x-axis. *Note that the supra-threshold response times of the JND measurements (green and red) were similar in profile to the JNDs (thick grey), but not to the response times of the speeded discrimination task (thin black). Moreover, in the second group response times were much higher in the speeded discrimination task than in the JND measurements.*

Comparison with JNDs and speeded discrimination (Tab. S8)

To test whether supra-threshold response times followed the pattern of JNDs, we calculated correlations between those response times (green and red curves, respectively) and the JNDs (thick grey curves in Figure S20). Table S8.a summarizes the results. Response times and JNDs were positively correlated across hues in both groups ($r(72) = 0.72$, $p < 0.001$, $R^2 = 52\%$, and $r(20) = 0.88$, $p < 0.001$, $R^2 = 77\%$). These correlations show that the response times for supra-threshold stimuli during JND measurements followed closely the JNDs that resulted from the adaptive staircase in this task.

In contrast, the patterns of supra-threshold response times in the JND measurements were different from those of the response times and error rates in the speeded discrimination task. Correlations across stimuli were calculated between the supra-threshold response times of the JND measurements on the one hand (green and red curves in Figure S20), and the response times and error-rates in the speeded discrimination task, on the other (thin black curves in Figure S20). As above, JNDs of the first group were interpolated for the 20 color pairs based on the original 72 JNDs.

Results are summarized in Table S8.b. The correlation between the supra-threshold response times in the JND measurements and the error-rates in the speeded-discrimination task were marginally significant for the second group ($r(20) = 0.42$, $p = 0.07$, $R^2 = 18\%$); but none of the other 3 measurements of speeded discrimination (response times for both groups, error rates for first group) shared any variance with the respective response times of the JND measurements (all $R^2 \approx 0\%$, min. $p = 0.75$).

These results show that the supra-threshold response times in the JND measurements showed a similar pattern across test-colors as the JNDs measured in this task, but not as the supra-threshold response times and error rates measured in the speeded discrimination task.

Consequently, those supra-threshold response times of the JND measurements show as few categorical patterns as the JNDs. In particular, the response times in the pink and green category, and maybe in blue, were above the boundary line for the first group (green curve in in Figure S20.a), as it was the case for JNDs (Witzel & Gegenfurtner, 2013). In the second group, the response times of pink, yellow, and green, but not orange and purple, showed a categorical pattern (red curve in in Figure S20.b). The absence of categorical patterns for several categories contradicts the idea that these patterns are specific to the categories, as predicted by category effects.

Finally, Figure S20 also highlights the particularity of the second group's response times in the speeded discrimination task as compared to those of the first group and those in the JND measurements. In the first group (Figure S20.a), the green curve and the thick grey curve lie almost upon each other. This illustrates that the overall size of re-

sponse times was similar for supra-threshold stimuli in the preliminary JND measurements and the speeded discrimination task. In contrast, response times of the second group strongly differed in size across the two kinds of tasks. The speeded discrimination task yielded much slower response times than the JND measurements in the second, but not in the first group.

1. Measure	2. Measure	n	r	p	R^2
a.) JNDs					
Pre-JNDs G1	JND-RTs G1	72	0.72	***	52%
Post-JNDs G2	JND-RTs G2	20	0.88	***	77%
b.) Speeded discrimination					
JND-RTs G1	Speed RT G1	20	-0.01	0.95	0%
JND-RTs G1	Speed ER G1	20	0.07	0.75	0%
JND-RTs G2	Speed RT G2	20	0.04	0.85	0%
JND-RTs G2	Speed ER G2	20	0.42	°	18%

Table S8. RTs during JND measurements. Correlations between supra-threshold response times during JND measurements on the one hand, and JNDs (**a**) and performance in the speeded discrimination task (**b**) on the other hand.

Development during JND measurements

Development of response times across blocks

Figure 7 of the main article illustrates how supra-threshold response times develop across blocks and sessions of the JND measurements. Both groups increased their speed across sessions and blocks, yielding a significant negative correlation between supra-threshold response times and blocks in the first ($r(144) = -0.76$, $p < 0.001$, $R^2 = 58\%$) and second group ($r(60) = -0.54$, $p < 0.001$, $R^2 = 29\%$). These results show that the task of the JND measurements strongly reduced response speed for supra-threshold stimuli.

At the end of the preliminary JND measurements, the first group reached response times close to the ones they had in the speeded discrimination task (Figure 7.a). The first group's response times in the speeded discrimination task were still lower than at the end of the preliminary JND measurements. Hence, this group still increased its speed when starting the main experiment (horizontal green line in Figure 7.a). This indicates that the training effects of the JND measurements were transferred to the speeded discriminations task.

The second group (Figure 7.b) followed a similar learning curve throughout the JND measurements as the first group, but at a slightly higher level of response times. The fact that the second group did not reach the speed of the first group at the end of the JND measurements might be due to the fact that the post-hoc measurements involved less measurements. They required less than half of the blocks (60 blocks in 6 sessions) compared to the preliminary JND measurements (144 blocks in 12 sessions).

Category effects across sessions (Fig. S21-S22)

We examined whether category effects depend on training and experience with the JND measurements. In particular, we tested whether category effects existed at the beginning of the JND measurements, and disappeared over time. For this purpose, we considered both, the pattern of JNDs and the pattern of supra-threshold response times. We examined the development of these measures across sessions of repeated measurements.

First, we inspected the development of JNDs across repeated measurements (Figure S21). In particular, we contrasted the first and the last of the repeated measurements to disentangle potential category effects at the beginning from the variation of JNDs at the end of the JND measurements (lower row of Figure S21). There was no evidence for any traces of categorical patterns in these contrasts, neither for the first (Figure S21.b), nor for the second group (Figure S21.d).

Second, we examined the development of supra-threshold response times during JND measurements (Figure S22). Again, we contrasted the first and the last measurements (lower row of Figure 22). There were no traces of categorical patterns in the contrasts of the first group (Figure 22.b). In the second group, there were slightly stronger categorical patterns for green and purple in the first measurements compared to the last measurements (Figure 22.d).

In sum, at no point in time there were any consistent category effects for the first group. This result supports the idea that the JND measurements counteract category effects. In contrast, the second group showed some faint support for the idea that there were some traces of category effects in the supra-threshold response times at the beginning of the JND measurements as compared to the end of the JND measurements. This latter result is in line with the idea that the second group carried over category effects from the speeded discrimination task, and that these effects disappeared because the JND task counteracted category effects.

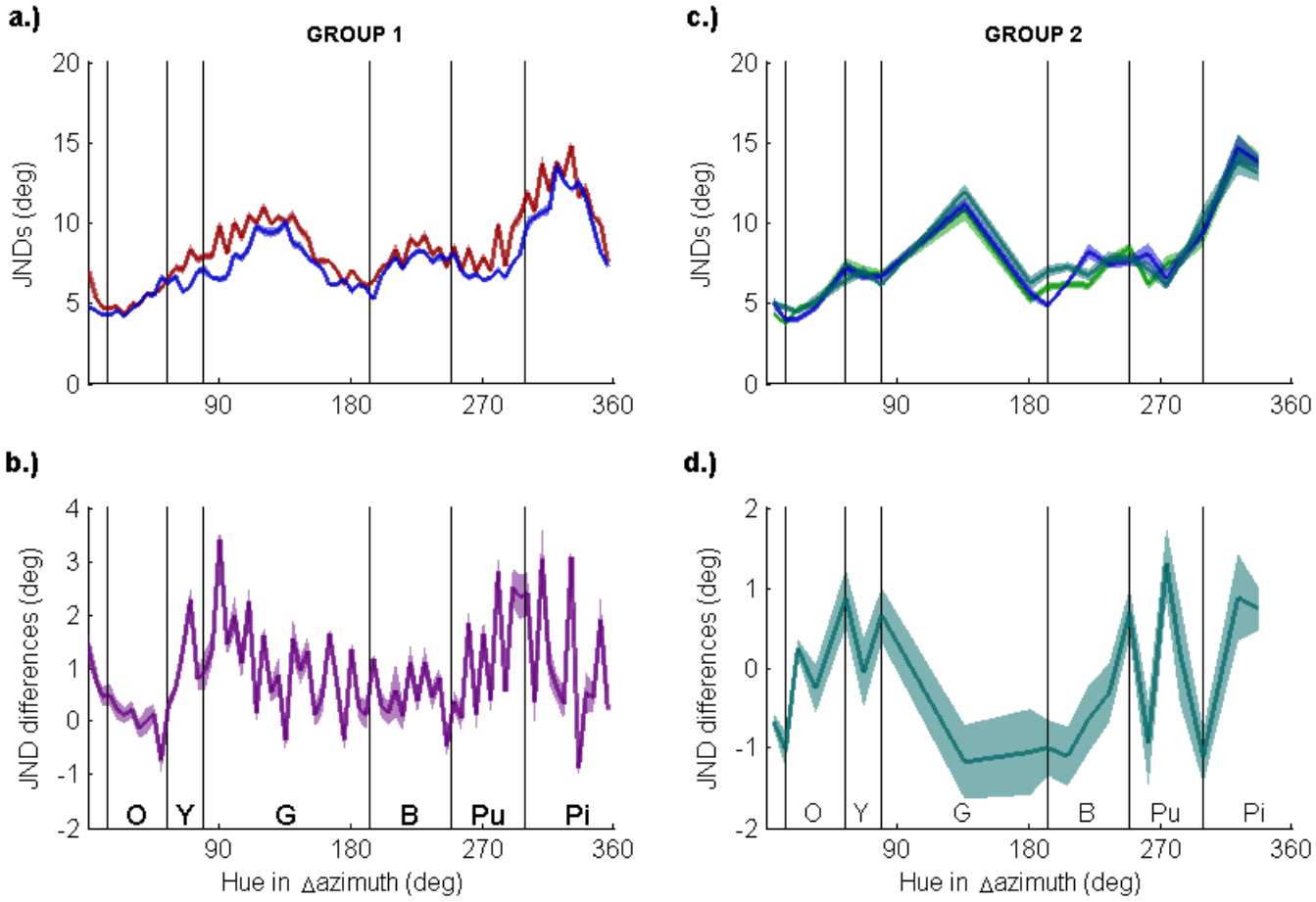

Figure S21. JNDs across sessions. Panel **a** shows the first (red) and the second (blue) of two repeated measurements in the preliminary JND measurements with the first group of participants. Panel **b** shows the differences between the first and the second measurements (i.e. difference between red and blue curve in panel a). Panel **c** illustrates the first (green), second (turquoise), and third (blue) measurements of JNDs in the post-hoc JND measurements with the second group. Panel **d** shows the difference between the first and the third measurement (i.e. between the green and the blue curves in panel c). Format as in Figure S20. *Note that the different measurements of JNDs were fairly stable across time, and there were neither categorical patterns in the differences between first and last JND measurements in the first group (panel b), nor in the second group (panel d).*

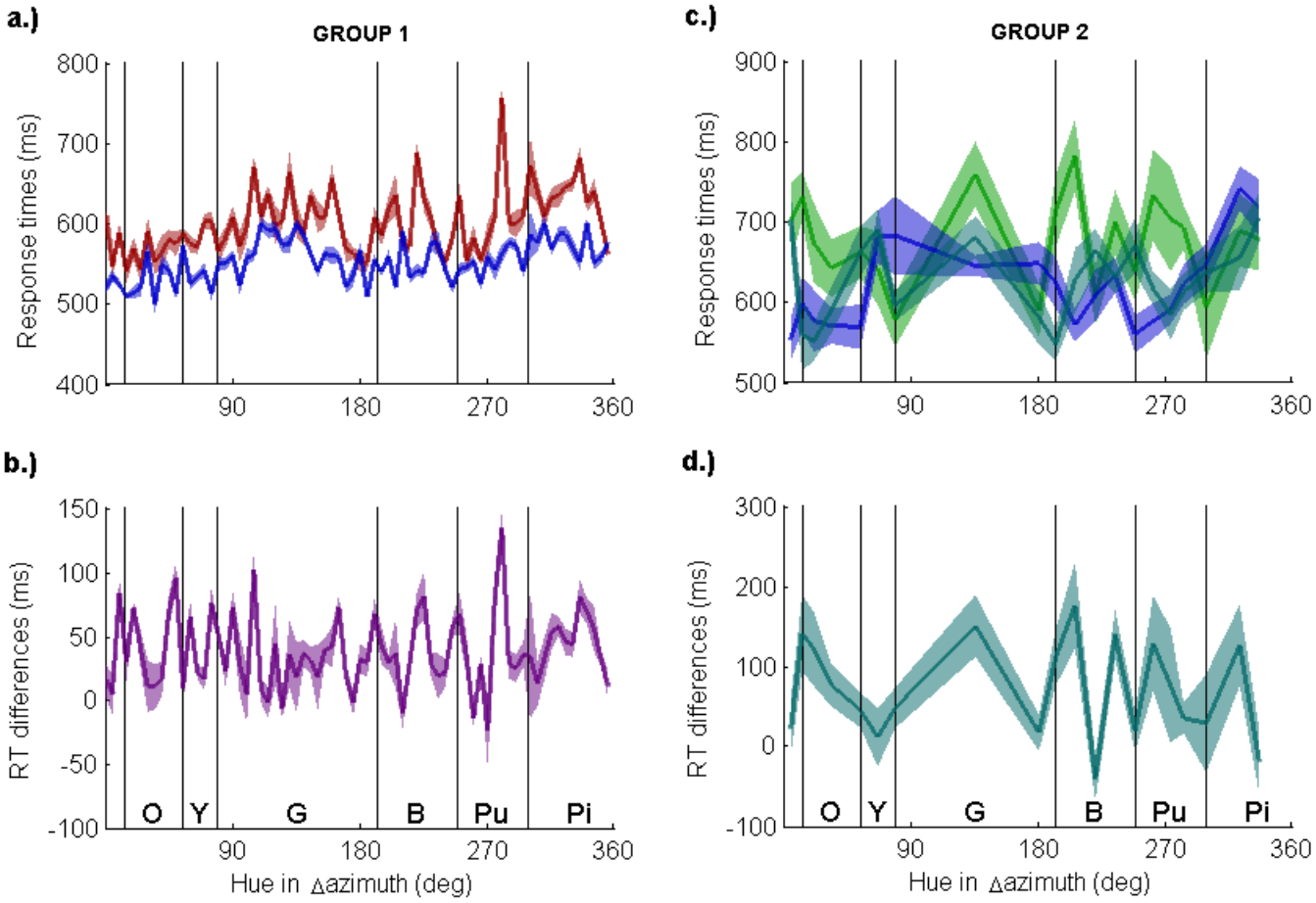

Figure S22. Supra-threshold response times across sessions of JND measurements. Panel **a** shows the response times for discriminating supra-threshold color differences in the first (red) and second (blue) of two repeated measurements in the preliminary JND measurements with the first group. Panel **b** depicts the differences betweeen the response times of the first (red curve in panel a) and the second measurements (blue curve in panel a). Panel **c** illustrates the supra-threshold response times during the first (green), second (turquoise), and third (blue) repeated measurements of the post-hoc measurements with the second group. Panel **d** shows the difference between the first (green curve in panel c) and the last (blue curve in panel c) post-hoc measurements. Apart from that, format as in Figure S21. *Note that there were no categorical patterns in the differences between first and last response time measurements in the first group (panel b), and some categorical patterns in those differences of the second group, but only for green and purple (panel d).*