Categorical color constancy for real surfaces

Maria Olkkonen Department of Psychology, University of Giessen, Giessen, Germany, & Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA

Christoph Witzel Department of Psychology, University of Giessen, Giessen, Germany

Thorsten Hansen Department of Psychology, University of Giessen, Giessen, Germany

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

In everyday experience, perceived colors of objects remain approximately constant under changes in illumination. This constancy is helpful for identifying objects across viewing conditions. Studies on color constancy often employ monitor simulations of illumination and reflectance changes. Real scenes, however, have features that might be important for color constancy but that are in general not captured by monitor displays. Here, we investigate categorical color constancy employing real surfaces and real illuminants in a rich viewing context. Observers sorted 450 Munsell samples into the 11 basic color categories under a daylight and four filtered daylight illuminants. We additionally manipulated illuminant cues from the local surround. Color constancy as quantified both with a classification consistency index and a standard color constancy index was high in both cue conditions. Observers generally classified colors with the same precision across different illuminants as across repetitions for the daylight illuminant. Moreover, the pattern of classification consistency in terms of stimulus hue, value, and chroma was similar when comparing different observers for the daylight illuminant and when comparing individual observers across different illuminants. We conclude that color categorization is robust under illuminant changes as well as across observers, thus potentially serving both object identification and communication.

Keywords: color appearance/constancy, color vision, color categorization, real surfaces, real scenes

Citation: 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.

Introduction

Did you ever have problems to judge whether a lemon is ripe or not? The light reflecting from objects to the eye varies substantially due to changes in lighting conditions—whether the objects are viewed under fluorescent light, sunlight, in shadow, or under a canopy in the forest. Regardless of these changes in illumination, we are usually able to successfully judge the ripeness of lemons, or more generally, the surface color of any object that we encounter. This ability to perceive constant surface colors despite the variability in the light signal is called color constancy.

Humans are able to discriminate thousands of colors (Linhares, Pinto, & Nascimento, 2008; Marín-Franch & Foster, 2010; Nickerson & Newhall, 1943; Pointer & Attridge, 1998), but colors are also readily classified into a few discrete categories (e.g., Berlin & Kay, 1969; Kay & Regier, 2003). Categorizing the colors of objects, as in the case of the ripening lemon green or yellow, might support color constancy in everyday situations together with lower level constancy mechanisms (Jameson & Hurvich, 1989; Smithson, 2005). In addition, color categories are useful when communicating about colors with others. For communication to be effective, however, two things are important: that color categories remain roughly constant under illuminant changes, and that there be some agreement between individuals about the use of categories. Our main goal here is to characterize categorical color constancy, but we also aim to elucidate the relationship between color constancy and communication by comparing the consistency of categories across illuminants to consistency of categories across individual observers.

In color constancy experiments, observers are commonly asked to match the appearance of two targets embedded in different illumination contexts (asymmetric matching), or to match one target to an internal reference of gray (achromatic settings). It is not clear whether these kinds of tasks are optimal for measuring constancy; in natural viewing conditions, illuminant changes often cause the color appearance of a surface to change, for

doi: 10.1167/10.9.16 Received April 20, 2010; published December 16, 2010 ISSN 1534-7362

instance across a shadow boundary, without affecting our judgment of the reflectance of the surface (e.g., Reeves, Amano, & Foster, 2008; Zaidi & Bostic, 2008). In other words, identical color appearance is not necessary for correct identification. Some more recent studies have investigated color constancy with tasks that do not require observers to match the appearance of stimuli across contexts (e.g., Bramwell & Hurlbert, 1996; Craven & Foster, 1992; Foster, Amano, & Nascimento, 2006; Zaidi & Bostic, 2008). Foster et al. (2006) adopted an operational approach to study color constancy in natural images. They found observers to be able to reliably discriminate between changes in the illuminant and changes in the reflectance of a test surface embedded in the scene; color constancy indices calculated from discrimination performance varied between 0.69 and 0.97. Zaidi and Bostic (2008) studied color constancy in real scenes with a forced-choice paradigm, where observers had to indicate which of four objects, placed in two contexts with different illuminants, was different from the other three. Zaidi and Bostic found that observers were often good at making this judgment, but that they made some systematic errors that could be described with a suboptimal similarity-based strategy. For the present study, we chose color classification as an alternative for matching, following a number of studies that have used it successfully to measure chromatic adaptation and color constancy (e.g., Amano & Foster, 2008; Chichilnisky & Wandell, 1999; Hansen, Walter, & Gegenfurtner, 2007; Olkkonen, Hansen, & Gegenfurtner, 2009; Smithson & Zaidi, 2004; Speigle & Brainard, 1996; Troost & de Weert, 1991; Uchikawa, Uchikawa, & Boynton, 1989).

One advantage of the color classification method is that it allows us to investigate the constancy of a large sample of both achromatic and chromatic surfaces. By using a collection of Munsell chips with varying hues, chromas, and values and by asking the observers to select prototypes among the chips in addition to the classification task, we can investigate whether constancy depends on the hue, chroma, or value of a particular chip, and whether prototypes are classified more consistently than other chips of the same chroma. The classification method also allows us to relate the consistency with which individual observers classify chips across illuminants to the consistency with which different observers classify the chips under the same illuminant.

Color constancy has been measured extensively with monitor simulations of flat surfaces and more recently of 3D scenes. Color constancy with these kinds of displays varies between low and relatively high (around 80%) depending on the exact task and richness of the display (e.g., Arend, Reeves, Schirillo, & Goldstein, 1991; Bäuml, 1994; Delahunt & Brainard, 2004; Lucassen & Walraven, 1996). The size of the illumination context also has a large effect on color constancy, at least for simple displays (Hansen et al., 2007; Murray, Daugirdiene, Vaitkevicius, Kulikowski, & Stanikunas, 2006; Rinner & Gegenfurtner, 2000). Moreover, cues from 3D scene geometry (Bloj, Kersten, & Hurlbert, 1999; Hedrich, Bloj, & Ruppertsberg, 2009), highlights (Snyder, Doerschner, & Maloney, 2005; Yang & Shevell, 2003), and stereo disparity (Werner, 2006; Yang & Shevell, 2002) appear to be used in estimating surface reflectance and illumination.

Using monitor displays is convenient because of the control they afford. Moreover, complex scenes containing physically accurate illuminant cues such as depth, shadows, and highlights can be readily simulated with the latest computer rendering techniques (e.g., Boyaci, Doerschner, Snyder, & Maloney, 2006). Still, some features of real scenes, such as depth along with the correct oculomotor cues, are hard to reproduce on conventional monitors (e.g., Hoffman, Girshick, Akeley, & Banks, 2008). In an important series of papers, Brainard et al. studied color constancy in a nearly natural real setup (Brainard, 1998; Brainard, Brunt, & Speigle, 1997; Kraft & Brainard, 1999; Kraft, Maloney, & Brainard, 2002). In their investigations, observers matched the color of surfaces produced by superimposing a projected image on a real surface across illuminant contexts. The experimental rooms contained various objects that provided cues to the illuminant. In general, color constancy indices measured either with asymmetric matching (Brainard et al., 1997) or with achromatic settings (Brainard, 1998; Kraft & Brainard, 1999) were higher than what had previously been found with monitor simulations for the same type of task. In line with this, de Almeida, Fiadeiro, and Nascimento (2004) found high color constancy indices (0.81–0.93) derived from asymmetric matches in a real setup.

We recently reported color constancy measurements for simulated surfaces using a color naming method, showing that observers' color naming consistency across illuminants approached that of test–retest consistency for a single illuminant and correlated strongly with color naming consistency across observers (Olkkonen et al., 2009). An important motivation for the present study is to investigate whether these findings generalize to a real scene. It will be of interest to see whether the degree of constancy is comparable for real and for simulated surfaces, and perhaps more importantly, whether we find the same pattern of constancy across color space for real surfaces as for the simulated surfaces.

Granzier, Brenner, and Smeets (2009) recently studied categorical color constancy in a fully natural setting. Granzier et al. collected color names for six unsaturated paper samples in several indoor and outdoor locations and found observers to be relatively good (55% identification rate) at naming the papers across the different locations. However, their conclusions are limited by the small number of test samples. Here, we investigate categorical color constancy for a large collection of real surfaces in a natural indoor setting where the illuminant could be manipulated with colored filters. In particular, we ask what the limits of categorical color constancy are in a rich

environment, and how the degree of color constancy relates to the color classification consistency of individual observers over time and to the color classification consistency across different observers for the same illuminant.

Methods

Observers

Four naive observers (one male and three females, ages 20–26) and the author CW (male, age 29) took part in the experiment. All participants were native German speakers. All had normal color vision as tested with the Ishihara color plates and self-reported normal or corrected-to-normal visual acuity.

Stimuli

Three hundred twenty chromatic and 10 achromatic Munsell samples used in the World Color Survey (WCS; http://www.icsi.berkeley.edu/wcs/) served as stimuli. The chromatic WCS surfaces include 40 hues from all Munsell hue groups (with 2.5-unit distance) and vary in lightness from Munsell value 2 to 9. The WCS chips are always the most saturated papers of a given hue/value combination, chroma varying between 2 and 16. The 10 achromatic chips have values between 1.5 and 9.5. To more closely investigate the effect of saturation on color constancy, we chose an additional set of 120 unsaturated Munsell chips, consisting of 3 chips from each the 40 Munsell hue groups having Munsell chromas 2, 4, and 6 at Munsell value 5. A photograph of the chip collection is shown in Figure 2A.

The experiment was conducted in a room with office furniture and large windows on the southwest side of the room (Figure 1A). We collected data under unfiltered daylight illumination and four chromatic illuminants that were produced by covering the windows with Lee filters (http://www.leefilters.com). The filters were selected so that the chromaticities of the produced illuminants corresponded roughly to the cardinal axes of the DKL color space (red, bluish green, greenish yellow, and violet), which were the illuminants used in our experiment on color naming for simulated surfaces (Olkkonen et al., 2009).

All sessions were conducted during light hours. The illumination was measured off a white reference surface (Photo Research SR-2) before and after each session for observers IR and CW and before and after most sessions for the other observers. Detailed information about the illuminant chromaticities is provided in Supplementary Tables 2–6. Filter specifications along with the average CIE *xyY* values of the illuminants are provided in Table 1. Figure 1B shows the illuminant chromaticities for each

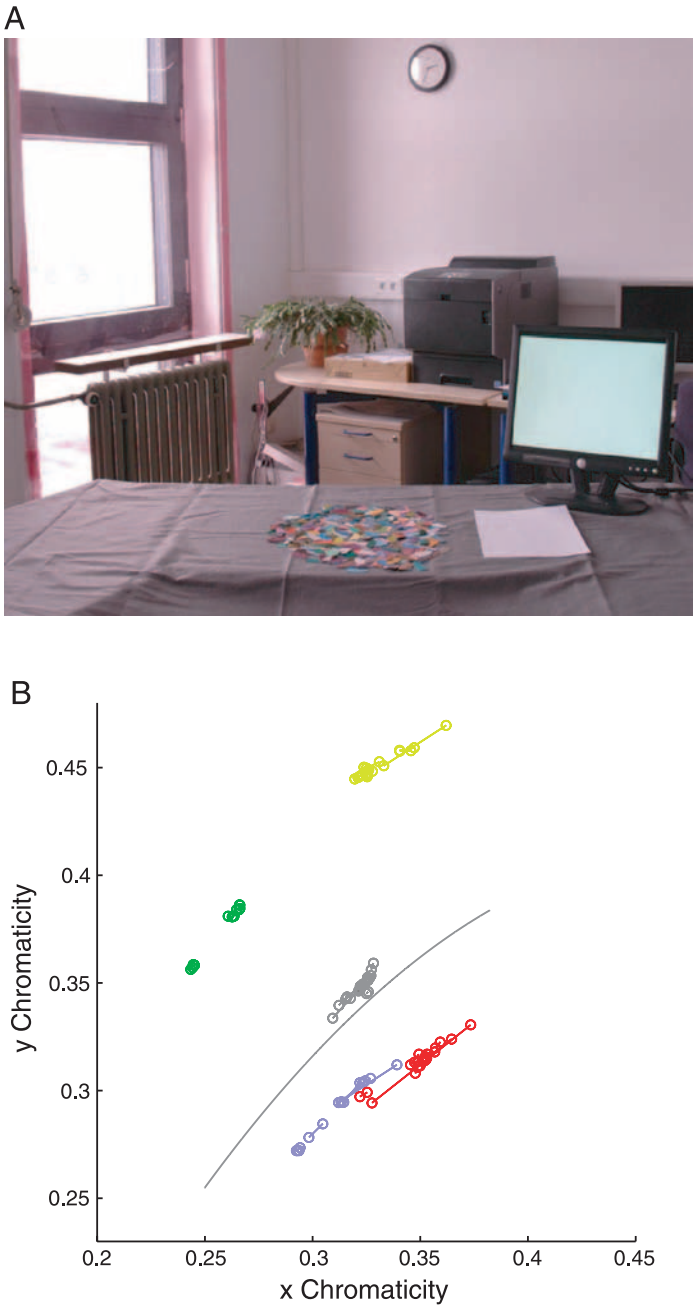


Figure 1. (A) A part of the experimental room with the red filters over the windows is shown with the chip collection on the gray cloth. The monitor and the white paper were not present during the experiments and are here only to indicate the reddish color appearance of the illuminant. The visible red borders of the filters at the edges of the window were covered during the experiments. (B) Judd–Vos corrected CIE chromaticities of the five experimental illuminants measured in each session are plotted on the *xy* plane. Symbol colors indicate the chromaticity of the illuminants. The two measurements made before and after each session are connected with lines. Different pairs of symbols show measurements in different sessions. The gray curve shows the daylight locus.

Illuminant	Filter	*x* mean (*SD*)	*y* mean (*SD*)	*Y* (cd/m^2) mean (*SD*)
Daylight	–	0.321 (0.005)	0.347 (0.005)	194 (137)
Red	35 Light pink	0.349 (0.012)	0.313 (0.009)	926 (2527)
Green	138 Pale green	0.256 (0.010)	0.372 (0.013)	280 (130)
Yellow	242 Lee 4300K	0.331 (0.01)	0.451 (0.007)	127 (146)
Violet	136 Pale lavender	0.310 (0.015)	0.290 (0.015)	206 (179)

Table 1. Filter specifications and the average Judd–Vos corrected CIE *xyY* values of the chip collection under each illuminant. The reported values are the means (standard deviations) across all sessions for a particular illuminant. Supplementary Tables 2–6 list the *xyY* values separately for each session and observer.

session where measurements were made. The variation in illuminant chromaticity within a given session was generally not large except for one session for the reddish illuminant for observer HB.

We used reflectance spectra provided by the University of Joensuu Color Group (http://spectral.joensuu.fi/) for calculating Judd–Vos corrected CIE *xyY* values of the chips under each illuminant. In order to see how well the calculated chromaticities matched the actual chromaticities under the illuminants, we measured the chromaticities directly from 10 saturated chips under the daylight and the yellowish illumination. There was a small difference between the measured and the computed chromaticities, as shown in Supplementary Figure 1. The Euclidean distance in the CIE *xy* plane between the measured and the computed chromaticities ranged between 0.009 and 0.03 for the daylight illuminant and between 0.003 and 0.03 for the yellowish illuminant. As our data analyses focus on comparing classification performance across conditions, this difference should not bear consequences on the interpretation of the results as any systematic shifts would be present for all illuminants. It should be noted, however, that the chromaticities depicted in Figure 2B and listed in Table 1 are based on the computed *xyY* values and are thus approximate.

Procedure

The experiment was run under two cue conditions. In the full-cue condition, the chips were laid out on a table on a medium gray cloth with Judd–Vos corrected CIE chromaticities $x = 0.33$, $y = 0.35$ under average daylight illumination. The chips were approximately square with an extension in the bottom part; maximum dimensions were 2 cm × 4 cm or 2.3 × 4.6 degrees of visual angle at a distance of 50 cm. The visible portion of the cloth was 148.5 cm × 61 cm or 112 × 63 degrees. In this condition, some of the chips had a higher luminance, and some had a lower luminance than the gray cloth.

In the reduced-cue condition, the chips were laid on a black cloth and observers wore black gloves to reduce cues to the illumination from local contrast between the chips and the background and between the chips and the hands. The visible portion of the black cloth was 200 cm × 61 cm or 127 × 63 degrees. In this condition, all chips reflected more light than the background.

In both conditions, all 450 chips were laid out on the cloth simultaneously in a random arrangement, and the observers' task was to sort the chips into eleven categories that correspond to the basic color terms (red, orange, yellow, green, brown, blue, purple, pink, white, gray, and black). The color names were given in German (rot, orange, gelb, grün, braun, blau, lila, rosa, weiß, grau, schwartz). Observers were not given any additional instructions on how to accomplish the sorting task.

In addition to sorting each chip, observers chose the best examples for each category, i.e., prototypical colors, among the chips in each condition.

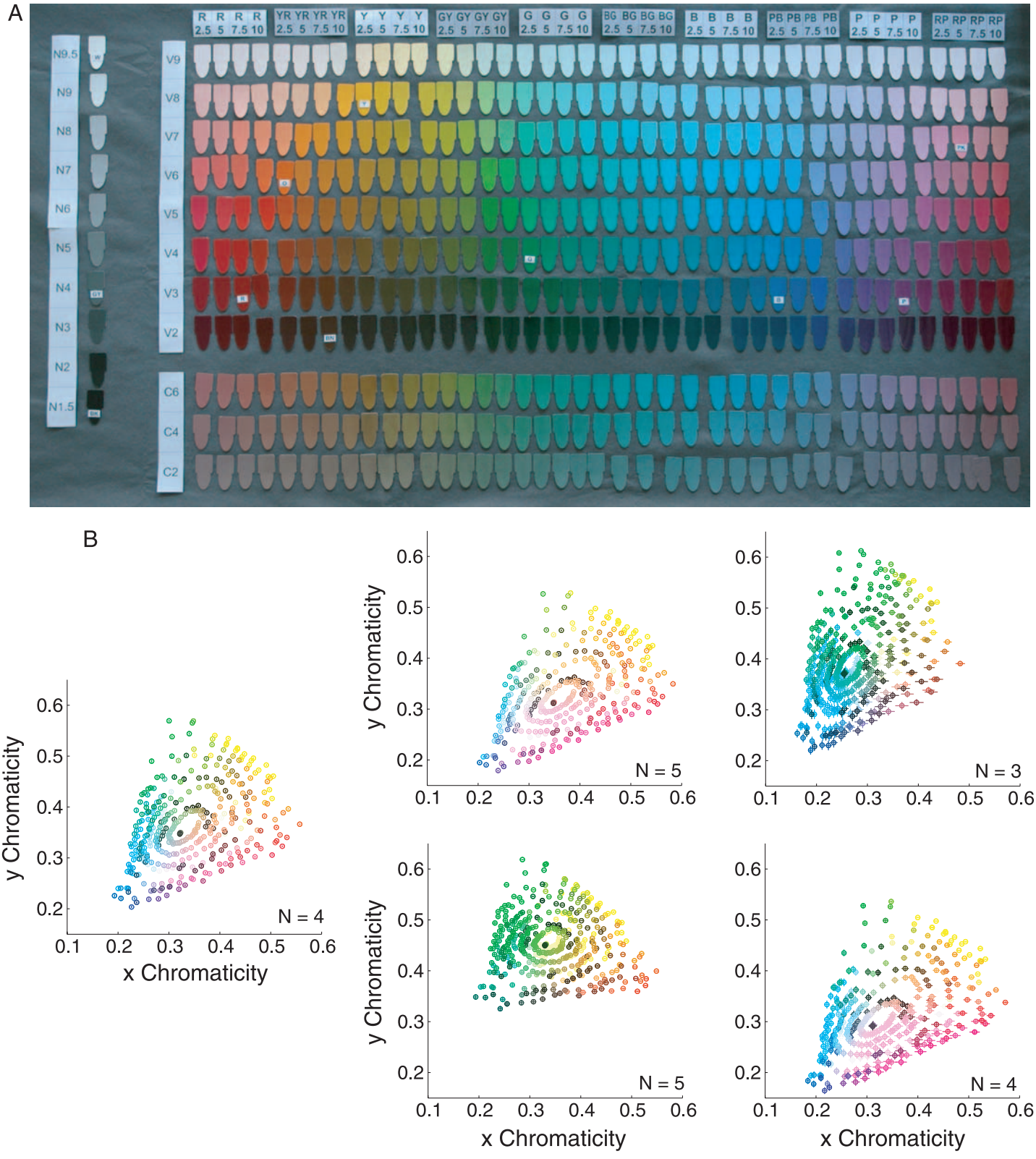
One session, taking about an hour, consisted of running the sorting task and the prototype selection once under one illuminant. Observers ran one session on any given day. Data for all observers were collected for a given illuminant condition in consecutive sessions, after which the filters on the windows were changed and the next condition was run. The order of the two cue conditions within any given illuminant condition was counterbalanced across observers. Supplementary Table 1 lists the session order for each observer.

Figure 2. (A) The WCS chips (8 top rows) and the unsaturated chips (3 bottom rows) are shown photographed under a daylight illuminant. The WCS chip collection is organized such that value increases from bottom to top and hue varies from left to right. The chroma for each chip is the maximum for that particular hue/value combination. In the unsaturated group, chroma decreases from top to bottom; all chips are at value 5. The ten achromatic chips are shown left from the WCS chips. (B) Average Judd–Vos corrected CIE *xy* chromaticities of the Munsell chips under each illuminant are shown for the daylight (far left) and for the reddish (top left), greenish (top right), yellowish (bottom left), and violet (bottom right) illuminants. The *XYZ* values for each chip were calculated from spectral measurements for the Munsell collection downloaded from the University of Joensuu Color Group Spectral Database. Illuminant spectra were measured in the experimental room for each filter. *N* in each panel indicates the number of available spectral measurements for each illuminant condition. Symbol colors indicate the color signal reflecting off the chips under each illuminant. Error bars show the variation in the color signals across repetitions of the same illuminant condition.

Data analysis

Color constancy

Color constancy was quantified with three different analyses. First, the proportion of same classifications between the second run of the daylight condition and each chromatic illuminant condition was calculated for each observer (*pairwise consistency*). For any given chip, this value could be 0 (different) or 1 (same), and the average across the whole stimulus collection describes



the overall amount of categorical color constancy for any given illuminant change from neutral. In addition, the proportion of same classifications between the first and second runs of the daylight condition was calculated for each observer to estimate test–retest color sorting reliability.

A second type of analysis sought to quantify the overall stability of categories across the five illuminant conditions (*overall consistency*). To this end, the frequency with which each Munsell chip was classified in the same category across all five illuminants was calculated for each observer (see Troost & de Weert, 1991). Specifically, the number of same classifications for each chip was calculated over the five illuminant conditions (range 0–5) and divided by the maximum number of same classifications (5). An index of 0 would mean that a chip was classified differently under all illuminations and 1 that a chip was classified the same under all illuminations. This index describes the overall degree of categorical color constancy for each chip. We also calculated classification consistency across observers for the daylight illuminant, which describes the degree of agreement between different observers' color categories.

Finally, the effect of each illuminant change from neutral on the achromatic point was quantified with a measure similar to a standard color constancy index (Equation 1). Achromatic points were defined as the centroids of the gray category in the Judd–Vos corrected CIE *xy* plane, i.e., the mean *xy* chromaticities of all chips named gray. Rather than to quantify the change in the achromatic point relative to the change in the illuminant, the physical change in stimulus chromaticities was used as the reference:

$$\mathrm{CI} = \frac{\mathbf{S}_c \cdot \mathbf{S}_p}{||\mathbf{S}_p||^2} \quad (1)$$

In Equation 1, vector $\mathbf{S}_c$ is the observed shift in the achromatic point from the neutral to a given test illuminant and vector $\mathbf{S}_p$ is the predicted shift of the achromatic point given perfect constancy. Projecting $\mathbf{S}_c$ to $\mathbf{S}_p$ gives the common component of the observed shift in the direction of the predicted shift, and the constancy index is derived by dividing the magnitude of the projection by the magnitude of the predicted shift.

We also calculated color constancy after Equation 1 for the category prototypes of each observer to get a comparison between color constancy for achromatic and chromatic samples.

Lower bound prediction

The fact that some color categories are larger than others might be expected to influence classification consistency across illuminants. In the case of a large category, such as green, even a large shift in the illuminant might not be sufficient to push the chromaticities of all stimuli to a different category. In other words, we expect some baseline consistency even in the absence of color constancy. To estimate the influence of this effect on color classification consistency, we calculated a lower bound estimate for consistency as follows. First, category boundaries were fitted to the color classification data under the daylight illuminant as described in Olkkonen et al. (2009). For boundary fitting, stimulus chromaticities were converted from CIE *XYZ* to CIE *L*a*b** coordinates (Wyszecki & Stiles, 1982). The CIE *L*a*b** space was convenient for boundary fitting because it represents color variation around an origin, which we defined as the chromaticity of the gray cloth under neutral daylight. The white point of the space was the same for all illuminant conditions. The fitting was done for each observer and each Munsell value separately. After fitting the boundaries, the light signal reflecting from each chip under each average chromatic illuminant was calculated in *L*a*b** coordinates. The light signals under each chromatic illuminant were categorized based on the category boundaries for the daylight illuminant. Finally, a consistency index was calculated based on the simulated classifications as described in the Color constancy section.

We had full illuminant measurements for two observers (IR and CW), for whom we were able to run the above analysis individually. Figure 3 shows the lower bound predictions calculated from the individual illuminant data for observers IR and CW, as well as from illuminant data averaged over all measurements. The patterns in the three curves are reasonably similar, and so we will use

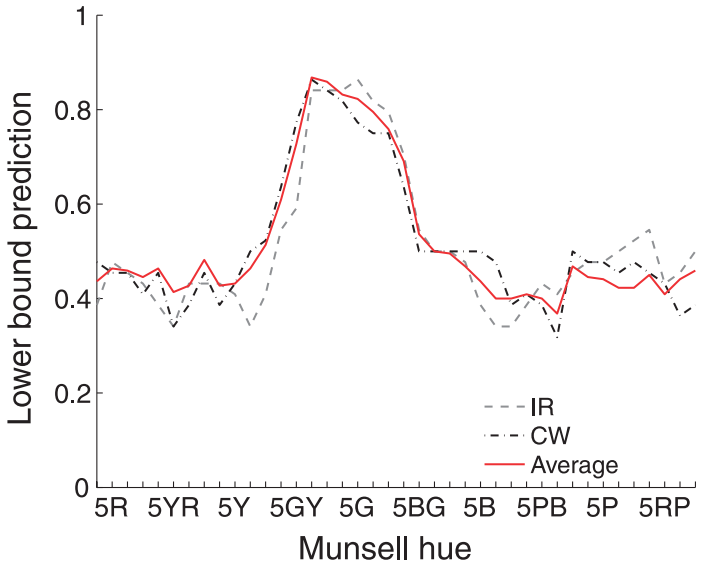


Figure 3. Lower bound predictions for individual and average illuminants as a function of Munsell hue. The lower bound prediction calculated from the illuminant spectra in each session is shown for observer IR with the gray dashed curve and for CW with the black dot-dash curve. The red solid curve shows the lower bound prediction calculated from average illuminant spectra.

individual predictions for IR and CW and the average predictions for the other three observers.

A lower bound estimate for observer consistency was calculated in a similar manner. As there are no stimulus chromaticity changes across observers, we modeled the effect of category size by rotating the category boundaries in the daylight condition by a random amount chosen uniformly between −60 and +60 degrees separately for each observer, after which the estimated observer consistency index was calculated based on the rotated boundaries. The rationale was that the largest categories would be most immune to the rotation and would thus give us an estimate of the effect of category size for the classification consistency across observers. The simulation was repeated 100 times, and the mean of the simulated indices was taken as the lower bound prediction for observer consistency.

Results

Classification consistency across illuminants

Observers classified colors in a rather consistent manner across different illuminants. Figure 4 shows the classification data for observer IR under all five illuminants in the full-cue conditions for Munsell chips at value 5. The data from the two sessions under daylight illumination are shown in the top row, and the data from the filter conditions are shown in the middle and bottom rows. This observer changed her classifying strategy for some categories between the first and second runs of the daylight condition. The categories remained more or less stable from the second baseline run to the filter conditions. For the other observers (Supplementary Figures 2–6), color categories remained similar across all conditions. Because of practice effects, we used the second run of the daylight condition as a baseline for subsequent color constancy calculations.

Figure 5 shows the color categories averaged over observers for the whole chromatic chip collection in the full-cue conditions. Category boundaries fitted to the data in the second baseline condition (Figure 5B) are shown in each panel with black lines. Comparing the data in each panel to the baseline boundaries shows that, overall, categories were rather stable across illuminants. The most salient changes were the enlargement of the green category between the baseline and the yellowish filter condition (Figure 5E), and the small decrease in the size of the pink category between baseline and the greenish filter condition (Figure 5D).

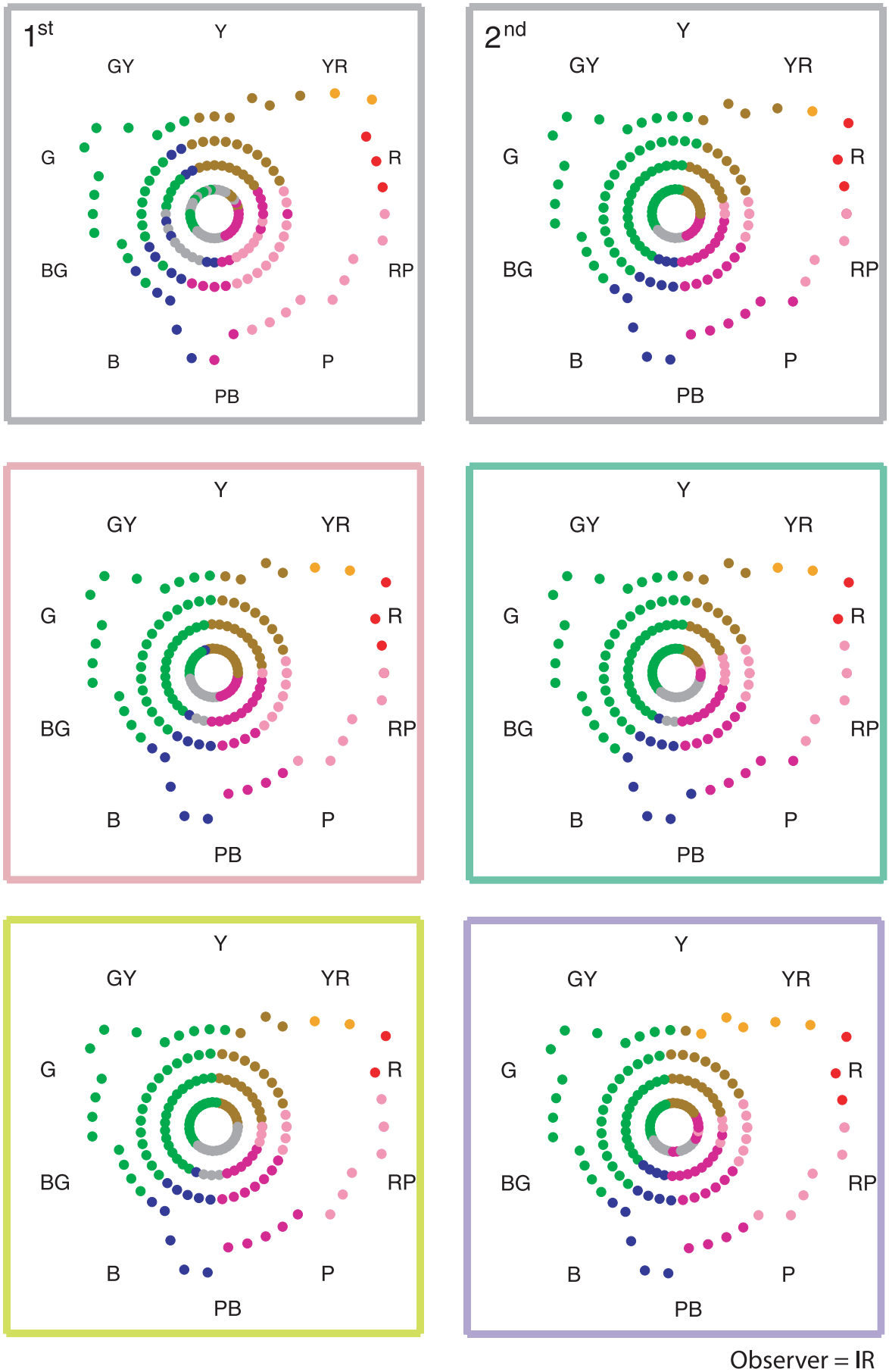
The small disks in each panel show the prototypical loci, aggregated over observers. The size of the disks indicates the frequency of each locus. Most loci tended to fall near the category centers, although there was some scatter especially for the pink category. Even though the loci remained rather constant under illuminant changes, the illuminant had an effect on the variability of the prototypes: under yellowish illumination (Figure 5E), for instance, the prototypes for red and green were the same for all observers, whereas under other illuminants these prototypes were more variable. Similarly, the prototype for pink was less variable under the violet illuminant (Figure 5F) than under the other illuminants.

In the reduced-cue conditions (Figure 6), categories were nearly as stable as in the full-cue conditions. Again, the most salient changes were the enlargement of the green category from daylight to the yellowish (Figure 6E) and to the greenish (Figure 6D) illuminants. As in the full-cue conditions, prototypes were generally located close to the category centers, with the exception of the red prototype, which was close to the red–brown boundary in all conditions. In addition, the prototype loci remained rather stable under illuminant changes. The constancy of prototypes is addressed further in the Color constancy for the prototypes section.

Similar figures for individual observers are presented in Supplementary Figures 2–11. The most salient effects for the individual observers were the enlargement of the green category under the yellowish illuminant (see, e.g., OX, AI, HB) and the enlargement of the purple category under the violet illuminant (see, e.g., OX, HB, CW). The effects were somewhat larger in the reduced-cue conditions particularly for OX, AI, HB, and CW.

Figures 5 and 6 show that both test–retest consistency and consistency across illuminants were rather high and of comparable degree. This is summarized for individual observers in Figure 7, which shows classification consistency between the two baseline runs, as well as consistency between the second baseline and each filter condition. For each observer, the first bar shows test–retest consistency in the baseline condition; the other bars show the comparisons between the baseline and each of the chromatic illuminant conditions. IR classified 62% of all stimuli in the same category between the first and second runs of the daylight condition and on average 84% of the stimuli when the baseline was compared to the filter conditions. For observers OX and HB, classification consistency was best for the test–retest condition at 93% and 97%, respectively, and dropped to 76% and 80% for the illuminant change comparisons. For observers AI and CW, classification consistency was about the same for the test–retest comparison at 83% and 79% and the illuminant

Figure 4. Raw classification data from observer IR under daylight illumination (first and second runs, top row), under the reddish and greenish illumination (middle row) and under the yellowish and violet illumination (bottom row). Each symbol denotes one Munsell chip at value 5. Symbol colors indicate the color name given to each chip. Munsell hue varies concentrically and chroma varies radially.



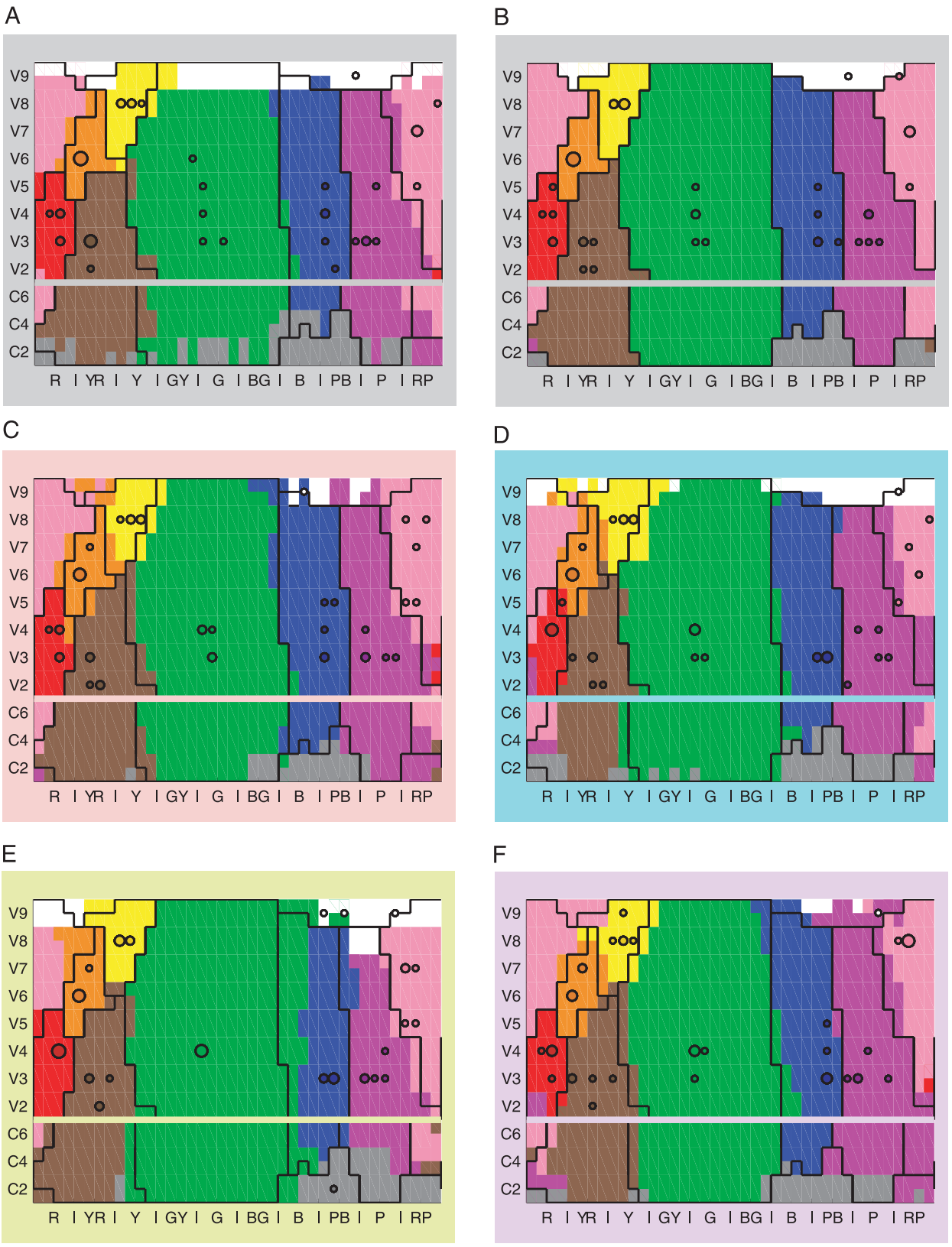


Figure 5. Average color categories in the full-cue conditions. Color categories aggregated over observers are shown for the (A) first and (B) second runs of the daylight condition. Hue varies from left to right, and value increases from bottom to top for the WCS chips (8 top rows). For the unsaturated chips (three bottom rows), saturation increases from bottom to top. The small disks show the prototypical loci. The size of the disks corresponds to the frequency with which each locus was selected. The black lines show category boundaries fitted to the second baseline condition. Aggregated color categories are shown for the (C) reddish, (D) greenish, (E) yellowish, and (F) violet filter conditions.

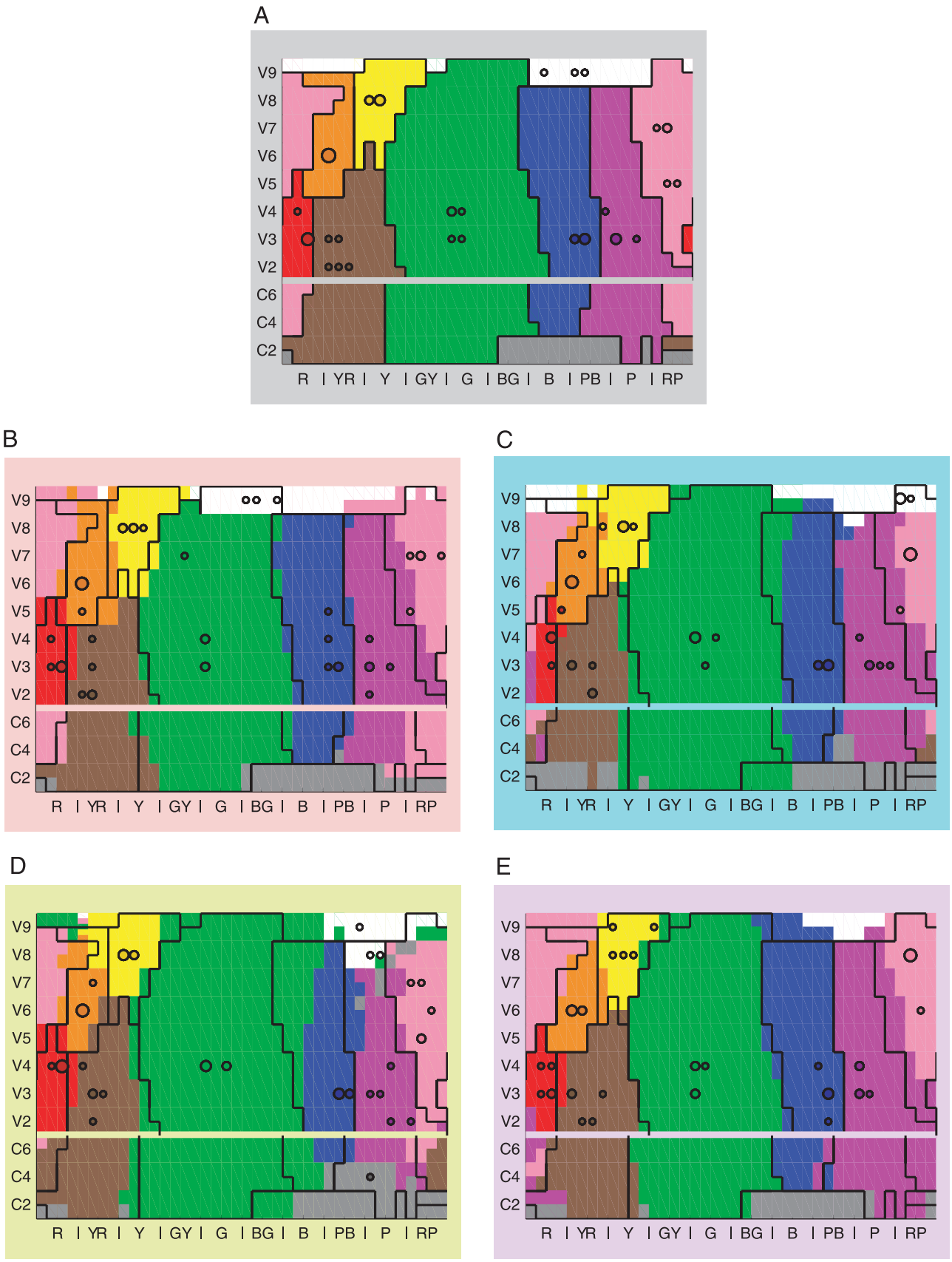


Figure 6. Aggregated color categories in the reduced-cue conditions. Details as in Figure 5.

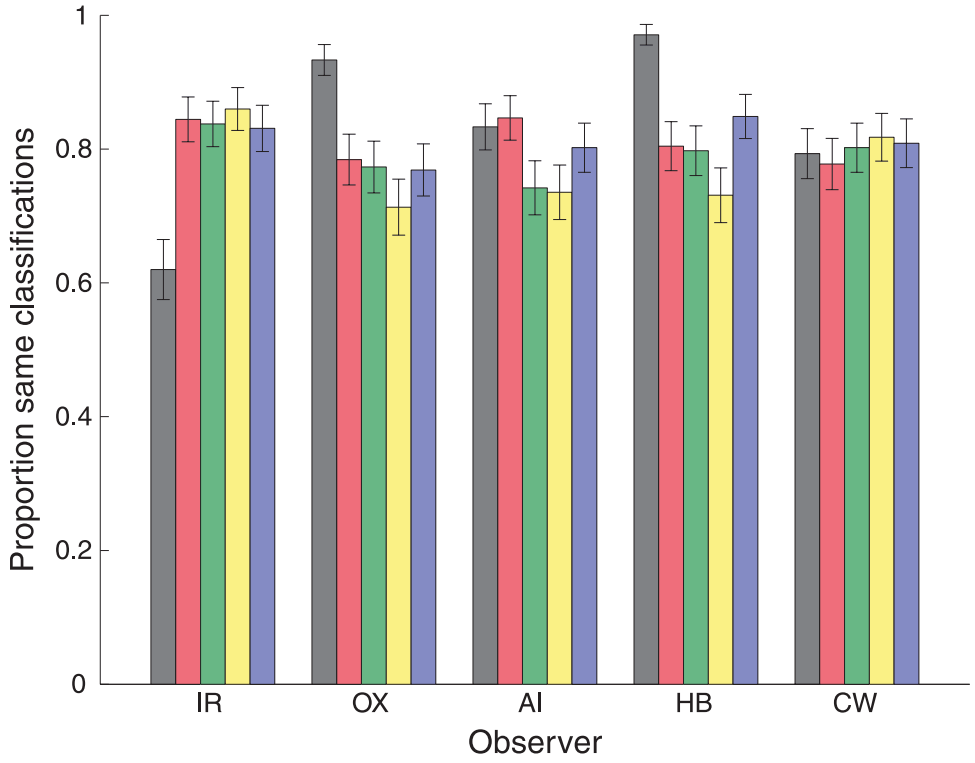


Figure 7. Proportion of same classifications pooled over stimuli for full-cue conditions. Each set of bars shows data for one observer. Each bar within a set shows the comparison of classification data between the daylight condition and a given other condition, from left to right, daylight 1–daylight 2, daylight 2–reddish, daylight 2–greenish, daylight 2–yellowish, daylight 2–violet. Error bars show the binomial proportion confidence intervals, where N = 450 (number of chips).

change comparisons at 79% and 80%. Binomial proportion confidence intervals (see error bars in Figure 7) indicate that the differences between the test–retest and the illuminant change conditions were significant for observers IR, OX, and HB but not for AI and CW. Note, however, that the direction of the effect was different for IR on the one hand, and for OX and HB on the other hand.

The fact that test–retest consistency was relatively low for IR compared to the other observers can be understood by looking at IR's raw classification data from the two runs of the daylight condition (see first row of Figure 4). The data from the first run were noisier than the data from the rest of the sessions, and the criteria for classifying chips green or purple changed from the first run to the second. In contrast, there were no such large changes for the other observers, which is clear from the raw data plots (Supplementary Figures 2–6) and from Figure 7.

Figure 8 illustrates the distribution of consistency over all observers and conditions (5 illuminants, 2 cue conditions). Numbers indicate relative consistency; values above 90% have been omitted to avoid clutter. Two things are clear from this figure. First, consistency was highest toward the centers of the categories. Second, prototypes tended to cluster in the central regions, perhaps with the exception of the pink and blue categories. Especially for the large green category, however, high consistency in the category center can be partly explained by the fact that the chromaticities of the stimuli close to category centers are bound not to shift as much under illuminant changes as the chromaticities of the chips close to boundaries, thus retaining their category membership under more extreme illumination changes. We will return to this issue in further analyses below.

Figure 9 illustrates the relative overall consistency of each category for both cue conditions. With the exception of the white category, consistency in both cue conditions was high. Overall classification consistency averaged over categories ranged from 75% to 82% for the full-cue condition (mean 80%) and from 75% to 83% for the reduced-cue condition (mean 78%) for the different observers. In a two-way repeated measures ANOVA with category and cue condition as factors, there was no statistically significant difference between cue conditions ($F(1, 4) = 2.2$, $p = 0.21$). In both cue conditions, consistency was highest for green, red, brown, and black, and lowest for pink and white. The overall difference between categories reached statistical significance ($F(10, 40) = 6.9$, $p < 0.001$).

We also analyzed consistency separately for Munsell hue, chroma, and value to see whether consistency depended systematically on any of these dimensions. In Figure 10A, overall consistency across illuminants averaged over observers is plotted as a function of hue for the full-cue condition (black circles) and for the reduced-cue

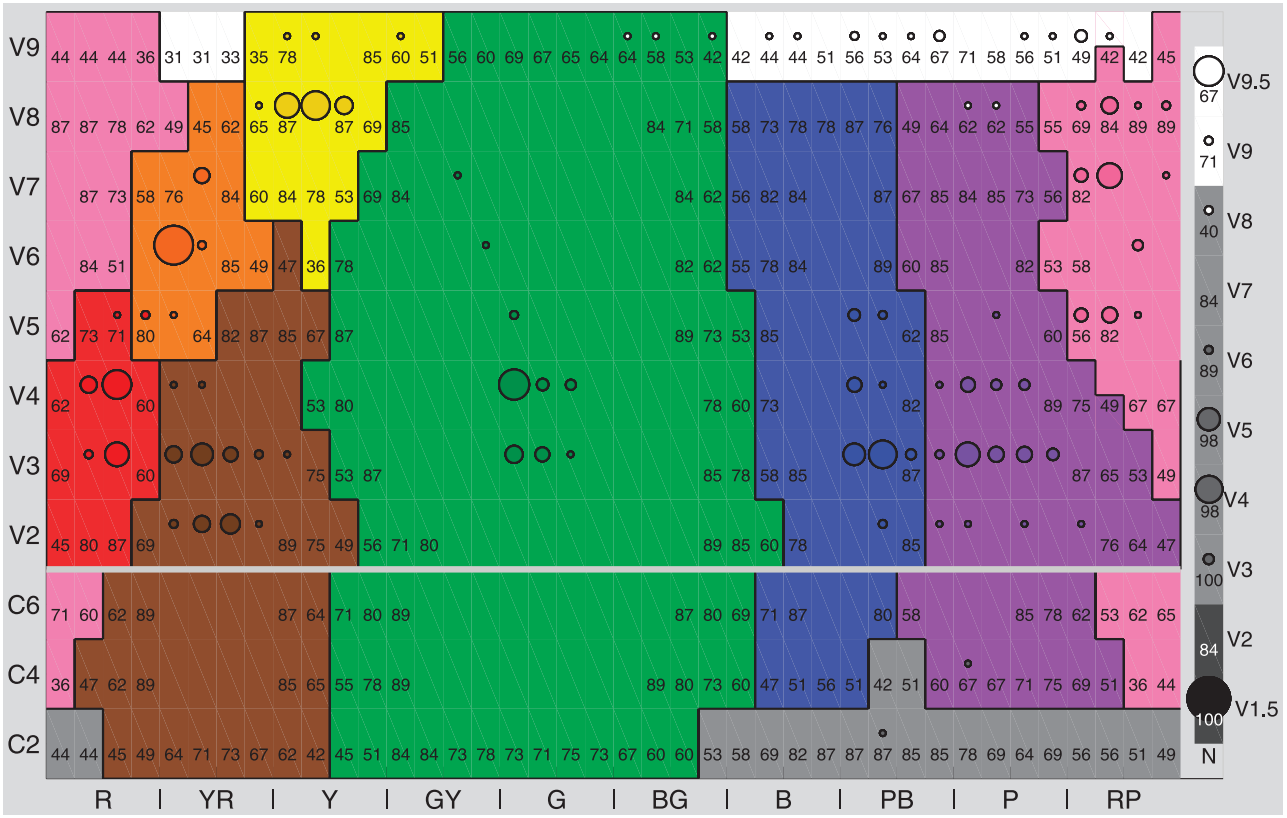


Figure 8. Stimulus consistency calculated over all observers, illuminations, and the two cue conditions. The chromatic chips are displayed in the same format as in Figures 5 and 6, and the achromatic chips are shown on the right. The small disks in each panel show the prototypical loci. The size of the disks corresponds to the frequency with which each locus was selected. The numbers show percent consistency; values above 90% have been omitted to avoid clutter.

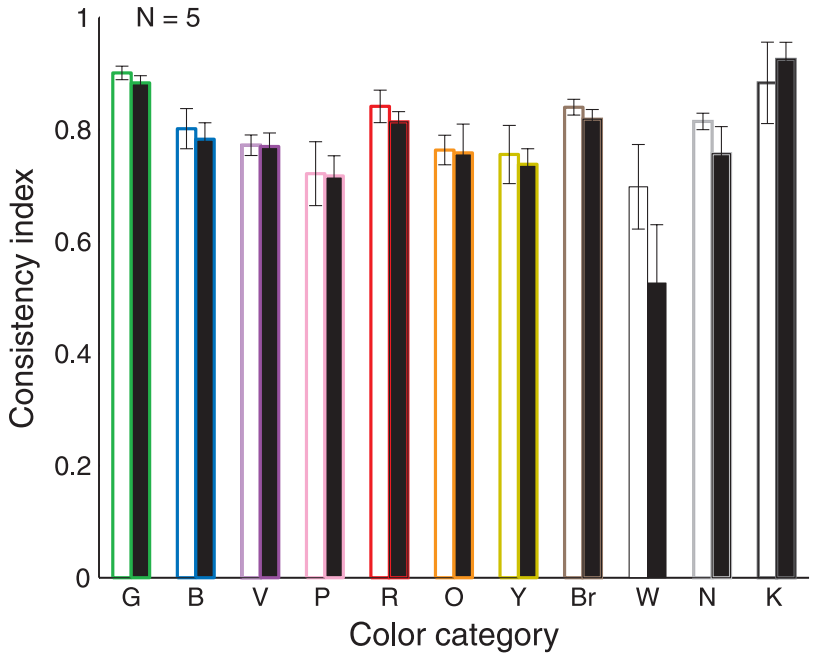


Figure 9. Overall classification consistency per category for full-cue conditions (white bars) and reduced-cue conditions (black bars). Error bars denote standard errors of the means across five observers. Color categories are indicated by the bar edge colors and denoted below the bars as follows: G (green), B (blue), V (purple), P (pink), R (red), O (orange), Y (yellow), Br (brown), W (white), N (gray), K (black).

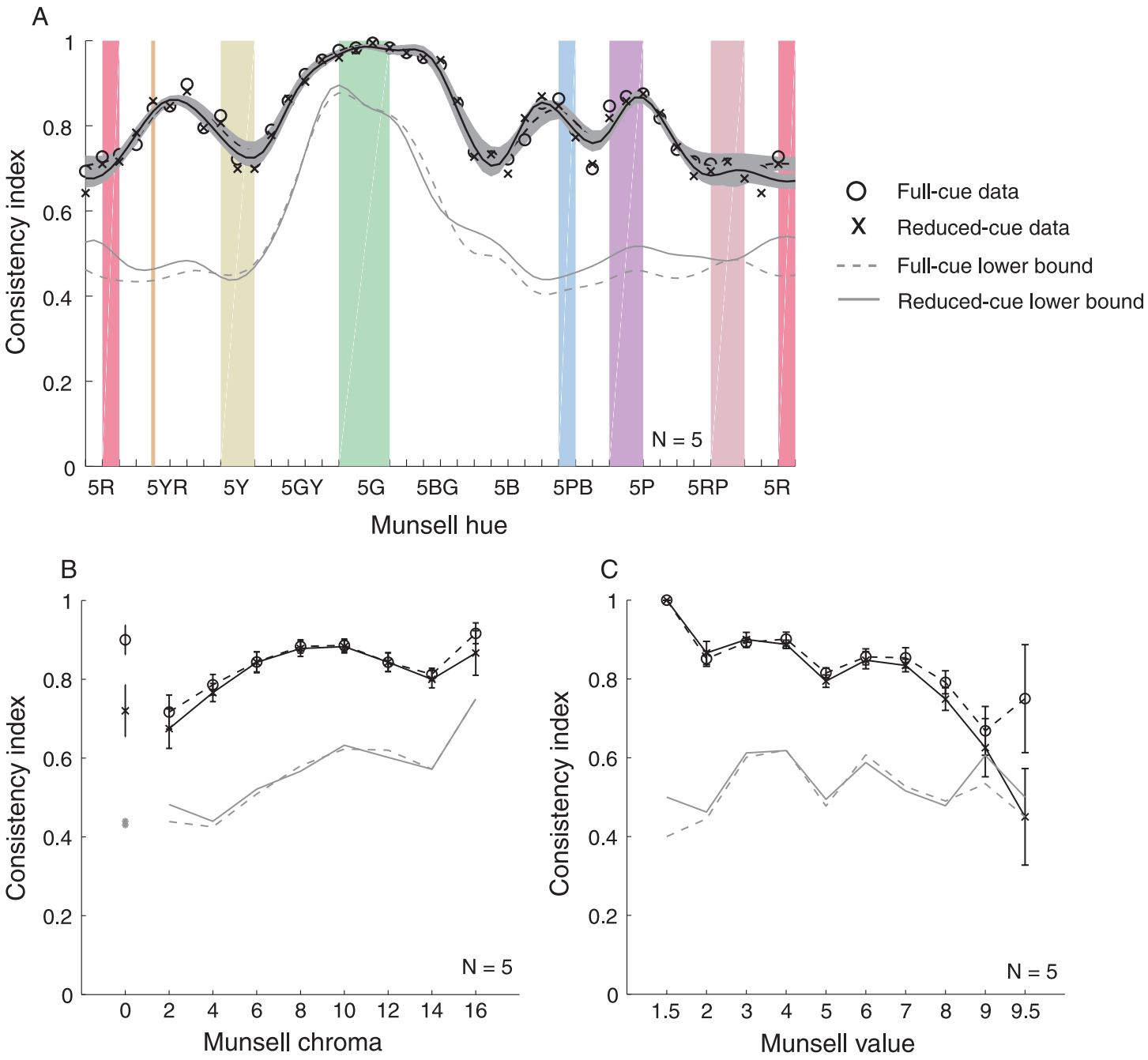


Figure 10. Across illuminants classification consistency as a function of stimulus variables. (A) Circles (full-cue condition) and crosses (reduced-cue condition) show overall classification consistency averaged over observers for each of the 40 Munsell hues. Indices have been pooled across stimulus value and chroma. The dashed and solid black curves drawn through the data points were derived by averaging values over two adjacent hues and interpolating between these averages. Shaded areas around the curves indicate the standard error of the mean across observers. Dashed and solid gray curves show the lower bound predictions for the full-cue and the reduced-cue conditions, respectively. Vertical bars indicate the range of prototypical hues chosen by the observers. (B) Naming consistency across stimulus chroma. The two data points for zero chroma at the left are for the achromatic chips. The rest of the data points are for the 440 chromatic chips, pooled over value and hue. Error bars indicate the standard error of the mean across observers. Other details as in (A). (C) Naming consistency across stimulus value pooled over stimulus hue and chroma. Details as in (B).

condition (black crosses). The gray dashed and solid curves plot the lower bound prediction for the full-cue and the reduced-cue conditions, respectively. The consistency maxima coincided with some of the prototypical hues (shown with shaded vertical bars), most notably with the green, blue, and purple prototypes. The lower bound prediction near green approached 0.9, however, indicating that in about 90% of the cases the stimuli would be

classified correctly without color constancy. For the other hues, the lower bound prediction was around 0.5. The comparison between the prediction and the data shows that classification consistency due to color constancy was particularly high for blue, purple, and orange and relatively low for pink and red.

Classification consistency increased with saturation for the chromatic chips (Figure 10B). The lower bound prediction indicates that the increase in consistency toward higher chromas was well predicted by the stability in stimulus chromaticities. In contrast, the high consistency for the neutral chips in the full-cue conditions, shown with the open symbol at Munsell chroma 0, was not reflected in the lower bound prediction. Analyzing consistency for the achromatic chips separately at each value revealed that consistency was only particularly good for the medium to dark achromatic chips; indices for achromatic chips at or below value 6 varied between 90% and 100% (mean 97%), whereas average index for achromatic chips above value 6 was 73%. In comparison, average consistency for chromatic chips was around 80%.

Figure 10C shows that for the whole chip collection, classification consistency tended to decrease with increasing value. This pattern was not obvious in the lower bound prediction, which was overall rather flat. The small trough in consistency for value 5 reflects the fact that the collection of unsaturated chips was selected from value level 5 and is thus overrepresented in the data point for that value.

Overall, consistency was similar in the full-cue and the reduced-cue conditions, which is indicated by the closeness of the dashed and solid black lines in all panels of Figure 10. The only exception to this were the achromatic chips, for which there was a large difference in consistency between the two cue conditions, shown in Figure 10B. Moreover, the pattern of consistency as a function of stimulus variables was virtually the same for both cue conditions.

Classification consistency across observers

Classification consistency across observers varied across stimulus hue in a manner similar to the consistency across illuminants (Figure 11A). The peaks in the consistency indices in both cue conditions coincided with some prototypes, particularly green, blue, purple, and orange. Consistency was lowest between the blue and the green prototypes and for the pink/red region. As in the case of the illuminant consistency index, the peak near green could be well explained by the lower bound prediction, but this was not as clearly the case for the other hues.

Figure 11B illustrates the fact that the variation in classification consistency as a function of hue was similar

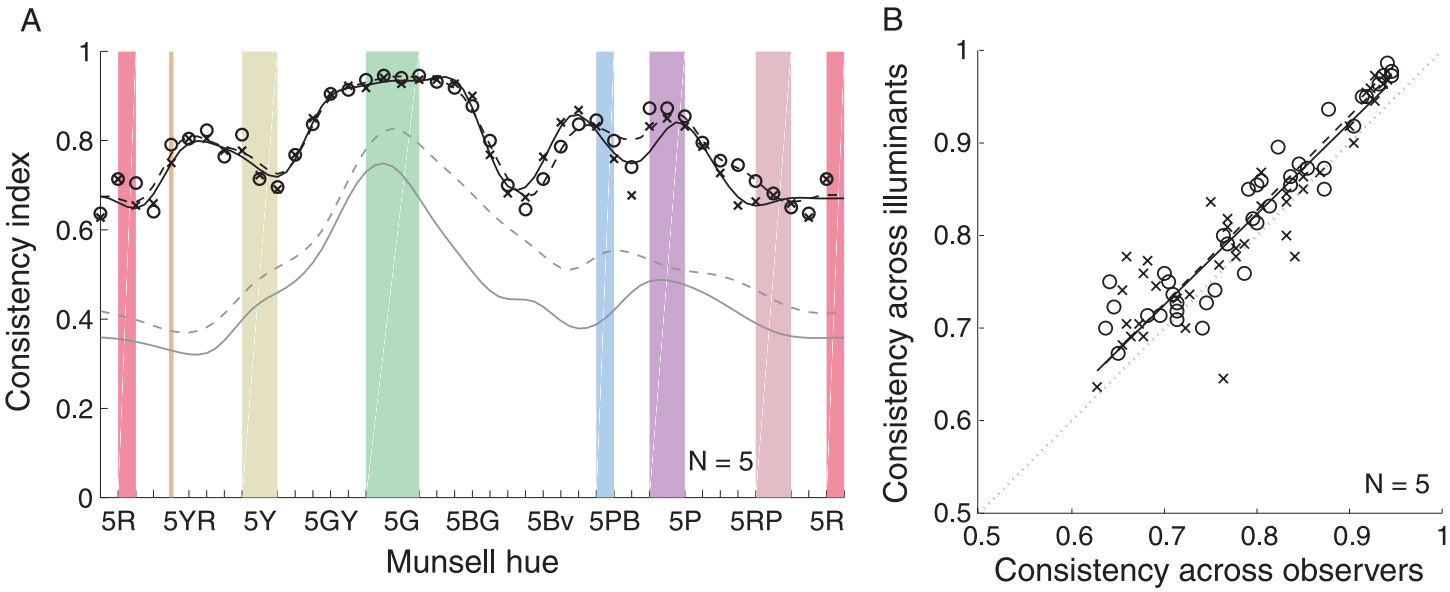


Figure 11. (A) Across observers classification consistency for the daylight illuminants is shown as a function of Munsell hue for the full-cue (circles) and reduced-cue (crosses) viewing conditions. The dashed (full-cue) and solid (reduced-cue) black curves drawn through the data points were derived by averaging values over two adjacent hues and interpolating between these averages. Gray dashed and solid curves show the lower bound predictions for the full-cue and reduced-cue conditions, respectively. Vertical shaded bars indicate the range of prototypical hues. (B) Classification consistency across illuminants is plotted against classification consistency across observers for the full-cue (circles) and the reduced-cue (crosses) viewing conditions. Each point indicates one Munsell hue, collapsed over chroma and value. The dashed and solid black lines correspond to the main common variance of the two indices (as represented by the first principal component) for the full-cue and the reduced-cue conditions, respectively. Gray dashed line indicates unity.

across illuminants and across observers. The two indices correlated strongly for both cue conditions (full-cue: $\rho = 0.95$, $p < 0.001$; reduced-cue: $\rho = 0.91$, $p < 0.001$). Partial correlations between the two indices, controlling for both the illuminant and the observer consistency lower bounds, were also high (full-cue: $\rho = 0.91$, $p < 0.001$; reduced-cue: $\rho = 0.83$, $p < 0.001$).

Color constancy for the achromatic point

To compare our classification data to previous color constancy experiments measuring achromatic settings, we quantified color constancy for the achromatic point with an index that relates the shift in the gray category centroid to the physical shift in stimulus chromaticities under a given illuminant change (Equation 1).

Color constancy indices ranged between 0.92 and 1.2 for the four illuminant changes from neutral for the full-cue conditions and between 0.88 and 1.02 for the reduced-cue conditions (Figure 12). Such high constancy indices are in line with data collected for monitor simulations in full-field viewing conditions (Hansen et al., 2007; Murray et al., 2006; Olkkonen et al., 2009; Rinner & Gegenfurtner, 2000) and for achromatic settings for real stimuli (Speigle & Brainard, 1999) but are higher than indices reported for asymmetric matching experiments in either simulated or real scenes (e.g., Arend et al., 1991; Brainard et al., 1997).

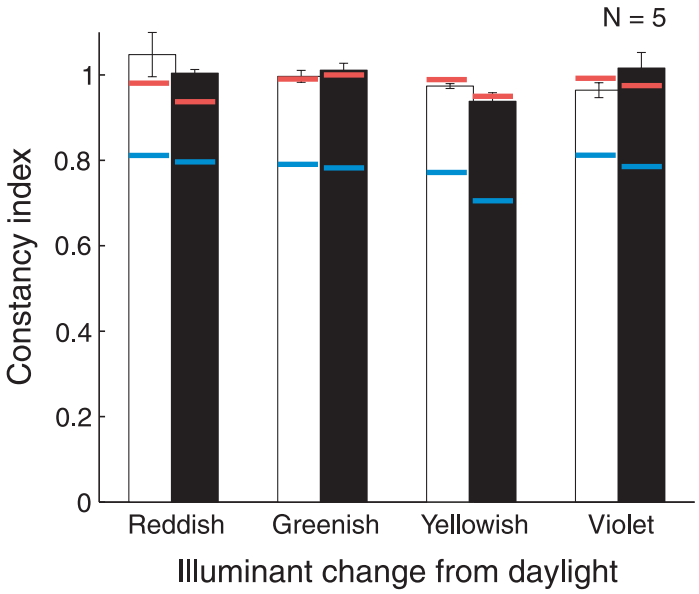


Figure 12. Average color constancy indices for the gray category centroid for the full-cue (white bars) and the reduced-cue (black bars) stimuli. For comparison, average pairwise classification consistency for the gray category is shown with the red lines, and average pairwise classification consistency for the whole chip collection is shown with the blue lines. Each set of bars is for a given illuminant change from baseline. Error bars denote one standard error of the mean over five observers.

The red lines drawn over the bars in Figure 12 show the proportion of same classifications between the daylight and each chromatic illuminant for the chips classified as gray. The blue lines show the proportion of same classifications aggregated over the whole stimulus collection. Classification consistency for the gray category gives roughly the same estimate for color constancy as does the color constancy index. On the other hand, classification consistency for the whole chip collection was lower at 80% and 78% for the full-cue and the reduced-cue conditions, respectively.

Color constancy for the prototypes

From Figure 12, it might appear that the chips classified as gray retained their perceived color, at least in a categorical sense, better than other chips in the stimulus set (compare the blue lines to the red lines and to the bars). To get some measure of color constancy for the chromatic samples comparable to the color constancy index for the achromatic point, we calculated color constancy indices for the chromatic category prototypes. Indices averaged over observers for each prototype and illuminant change from neutral are shown in Figure 13. Color constancy for the chromatic prototypes was on average close to 1.

From the high constancy indices alone, it is difficult to say whether category prototypes enjoy a higher consistency than other chips of the same Munsell chroma and value. However, comparing pairwise classification consistency for prototypes and non-prototypes of the same Munsell parameters suggests that constancy was indeed slightly higher for the prototypes. With the exception of pink and white, for which classification consistency was relatively low (70% and 40%, respectively), consistency was overall very high for prototypes (100% for green, purple, orange, brown, gray, and black, and 95% for red, yellow, and blue). In comparison, consistency for the rest of the chip collection was on average 82%. In a two-way repeated measures ANOVA with chip type (prototype/other) and chroma as factors, both the main effects of chroma ($F(3, 12) = 5.4$, $p = 0.01$) and chip type ($F(1, 4) = 28.8$, $p = 0.006$) were significant. There was no significant interaction between chroma and type ($p = 0.13$).

Discussion

We found a high degree of categorical color constancy for real surfaces across five broadband illuminants. This confirms previous findings with monitor simulations (Hansen et al., 2007; Olkkonen et al., 2009) and is in line with comparable results from hue scaling (Schultz, Doerschner, & Maloney, 2006), successive memory matching (Ling & Hurlbert, 2008), color selection (Hedrich et al.,

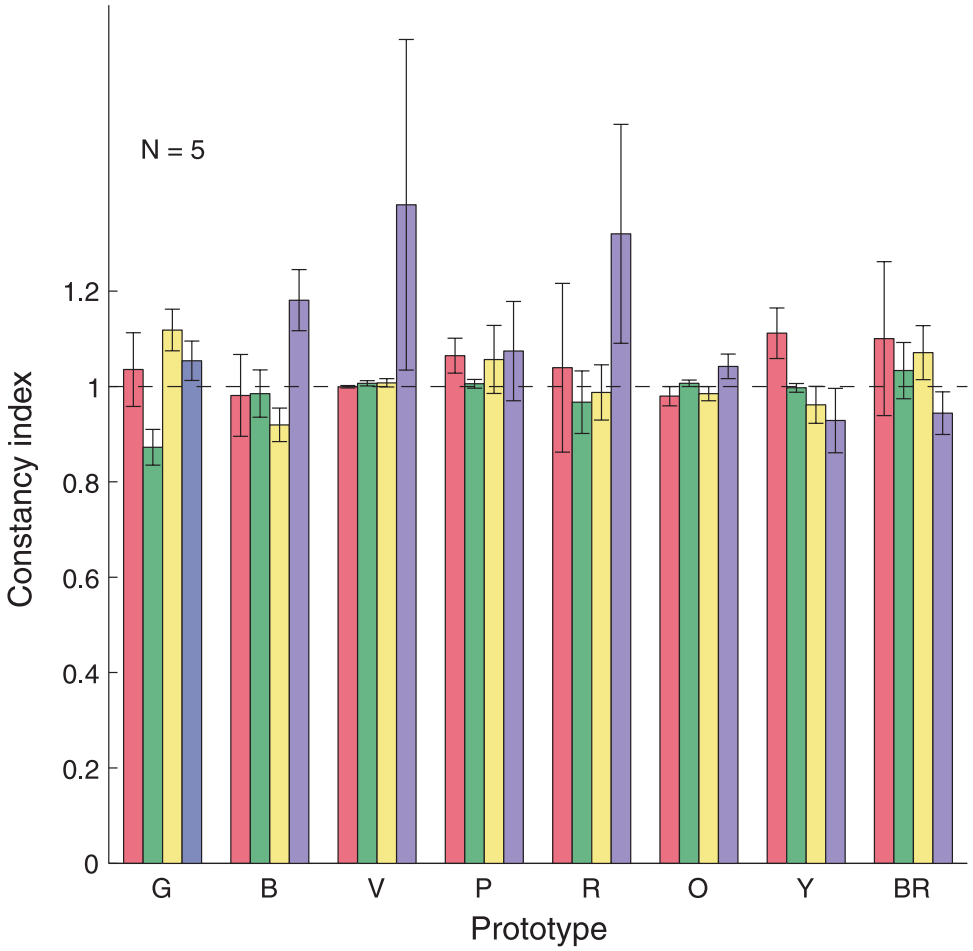


Figure 13. Color constancy indices for the prototypes in the full-cue conditions. Each set of bars is for one prototype, and the bars in each set stand for illuminant changes from daylight to the reddish, greenish, yellowish, and violet directions, from left to right. The notations of the prototypes are as in Figure 9. Error bars show one standard error of the mean over five observers. The dashed line indicates full constancy. Values over 1 indicate overcompensation for the illuminant change.

2009), and a preliminary report of categorical color constancy (Ling, Allen-Clarke, Vurro, & Hurlbert, 2008). We will discuss the main results in the following sections in relation to our previous work with simulated surfaces, as well as to other work on color constancy in simulated and real scenes.

Comparison to similar data for simulated surfaces

One motivation behind the present study was to follow up on experiments we reported recently on categorical color constancy for simulated surfaces (Olkkonen et al., 2009). The stimulus collections for the two experiments are shown in Figure 14A. Even though only 18 of the same chips were used in both studies, the range of chromaticities overlapped to a large extent. The chip collections in both experiments included chips from all Munsell hue groups, from value groups 4 to 7 and from chroma groups 2 to 12.

The chromaticities of the illuminants in the two experiments are shown in Figure 14B. We chose filters for the present study that would reproduce the chromaticities of the lamp illuminants from our previous study as closely as possible. As is clear from Figure 14B, the chromaticities of the filter illuminants, shown with open symbols, are shifted relative to the lamp illuminants, shown with crosses. The direction and magnitude of the offsets of the chromatic illuminants from the daylight illuminants, however, are comparable in the two cases.

Consistency across illuminants

Figure 15 shows a comparison between the present data and the data from Olkkonen et al. (2009) under full-cue conditions. In general, the two data sets agree remarkably well. Naming consistency across illuminants is higher for the real chips for some hues, but the pattern of consistency for hue (Figure 15A), chroma (Figure 15B), and value (Figure 15C) is similar for the levels that overlapped in the two experiments. In contrast to the classification data,

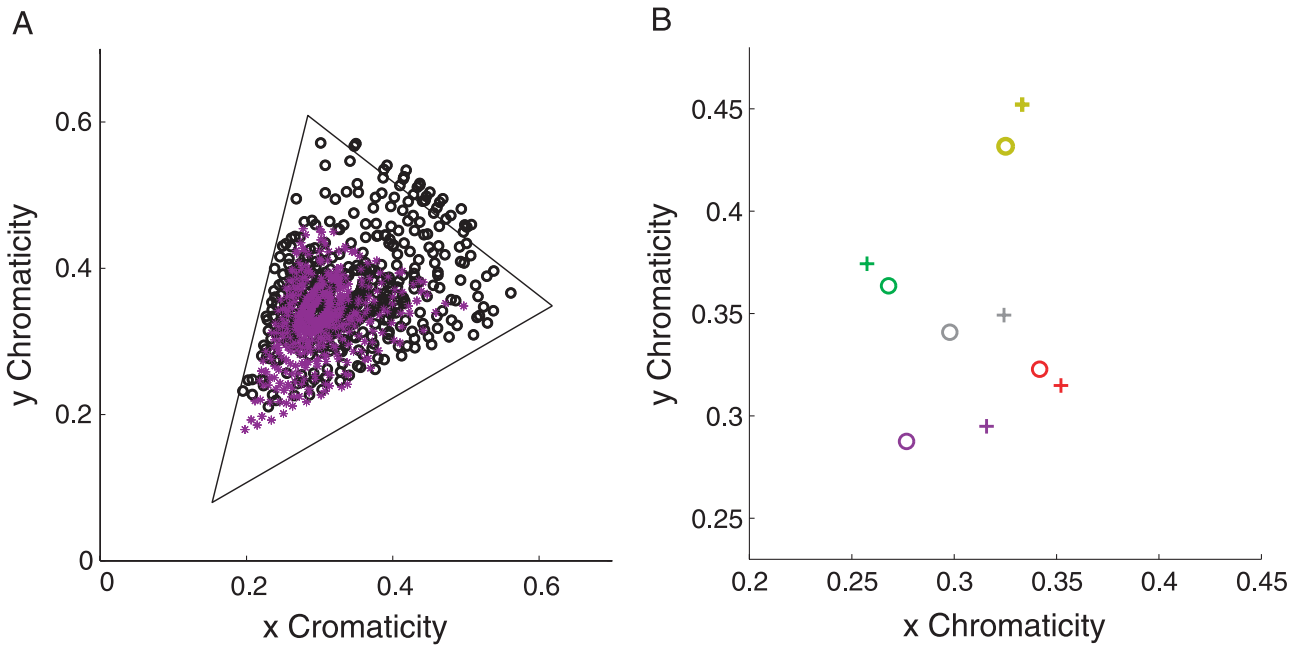


Figure 14. (A) Black open circles denote the Judd–Vos corrected CIE *xy* chromaticities of the real Munsell surfaces when viewed under a standard daylight. The purple asterisks denote the chromaticities of the simulated Munsell surfaces in Olkkonen et al. (2009) under a standard daylight. The black triangle indicates monitor gamut. (B) The CIE *xy* chromaticities of the real illuminants are shown with crosses and those of the lamp illuminants with open symbols. Symbol colors indicate the chromaticity of the illuminants. Note the different axis scales in (A) and (B).

the lower bound predictions are somewhat different for the two types of surface, reflecting the differences in the stimulus collections and illuminants employed in the two experiments. This further underlines the fact that classification performance is not solely determined by the shifts in the physical stimulus but also by category effects such as, for instance, the location of a stimulus within its category.

Based on both data sets, categorical color constancy relative to the baseline correction is highest near the orange, blue, and purple prototypes and relatively low at the pink and red prototypes. Consistency is also high for green, but there it is difficult to disentangle constancy from baseline stability because of a ceiling effect. It is worth noting that in the experiment with monitor simulations, the category "pink" was not available for observers. The lack of one basic color term might have conceivably caused a trough in the consistency function around that color, but as Figure 15A shows, consistency was equally low for the real as for the simulated surfaces around red. There is one caveat, however. One notable feature of the representation in Figure 15A is that consistency is pooled over value and chroma to highlight the pattern of consistency for stimulus hue. While the categories for blue, purple, and green do not depend on value in our data set, the categories for pink and red, as well as for brown, orange, and yellow partly overlap on the hue continuum (visible from Figure 5). In the case of red and pink, for instance, dark chips around Munsell red (R) are classified as red, whereas lighter chips are classified as pink. As can be seen from Figure 9, consistency for pink was lower than for red, which contributes to the consistency around the red prototype in Figure 15A.

When consistency is plotted two-dimensionally for hue and value, there is a visible tendency for consistency to peak near the prototypes (Figure 8). This implies that prototypes seem to be relatively stable across individuals and illuminations, when compared to boundaries. This observation has also been made for cross-language comparisons: boundaries vary considerably across languages, while prototypes are relatively stable (MacLaury, 1997; Regier, Kay, & Cook, 2005; Regier, Kay, & Khetarpal, 2007; Webster & Kay, 2005, 2007). This might indicate the special status of prototypes suggested by, e.g., Philipona and O'Regan (2006). As is also clear from Figure 8, however, prototypes tend to cluster near the category centers, and thus it is possible that the closeness to the category center is more important than prototypicality.

Consistency across observers

Observer consistency followed a similar pattern to illuminant consistency for both types of surface. Figure 15D plots classification consistency across illuminants against classification consistency across observers for the two experiments. There was a significant correlation between illuminant and observer consistency both for the real (black

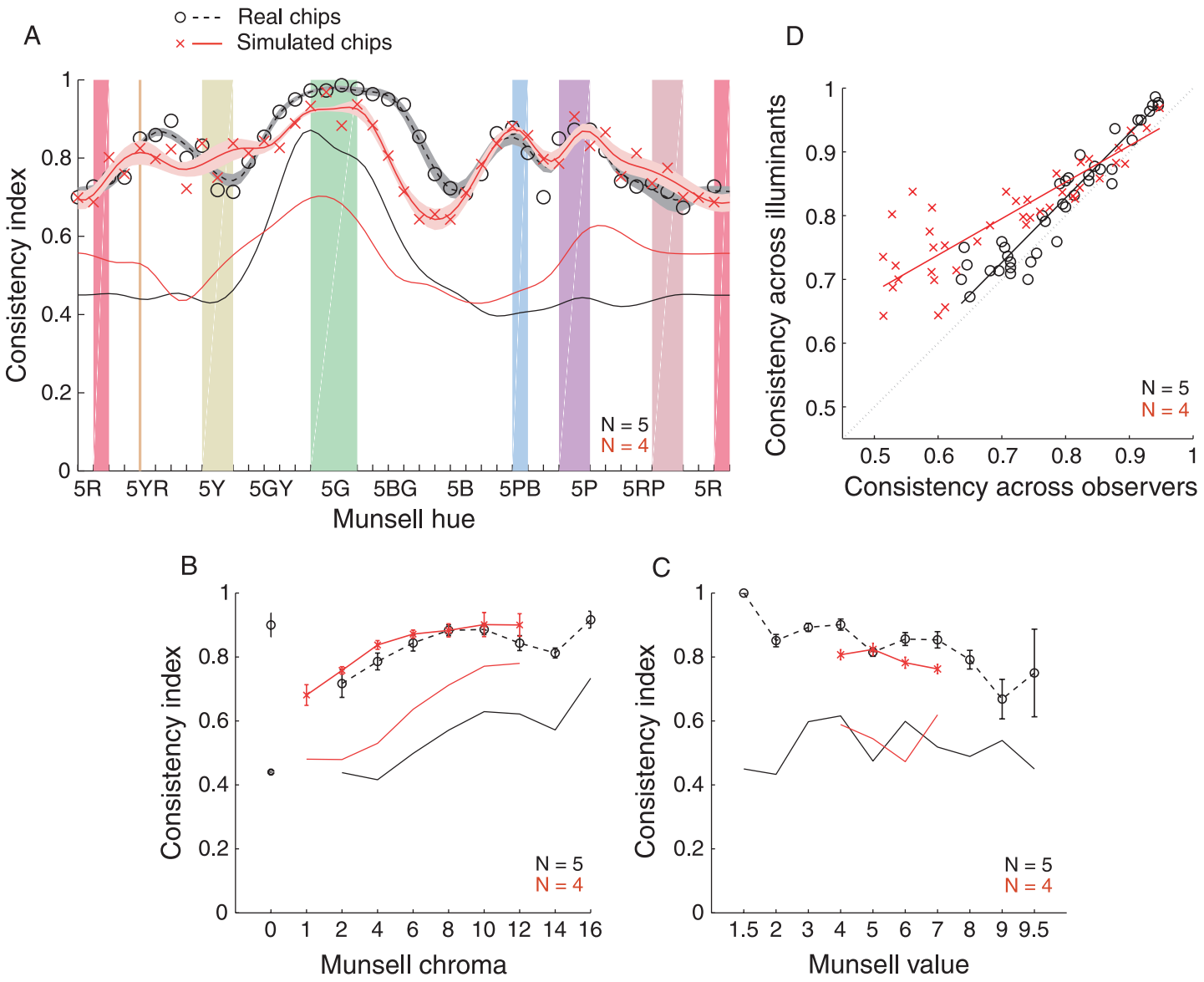


Figure 15. Comparison between color classification data for monitor simulations and real surfaces. (A–C) Classification consistency across illuminants for monitor simulations (red symbols and curves) and real surfaces (black symbols and curves), where the monitor data are replotted from Figures 5 and 7 in Olkkonen et al. (2009). Plotting conventions are otherwise as in Figure 10 of the present paper. (D) Classification consistency across illuminants is plotted against classification consistency across observers for monitor simulations (red crosses) and real surfaces (black circles). The black and red lines show the main common variance of the two indices (as represented by the first principal component) for the real and simulated surfaces, respectively. The gray dashed line indicates unity.

circles, $\rho = 0.95$, $p < 0.001$) and for the simulated (red crosses, $\rho = 0.85$, $p < 0.001$) surfaces. Partial correlations that controlled for the lower bound predictions were also high (real: $\rho = 0.91$, $p < 0.001$, simulated: $\rho = 0.81$, $p < 0.001$). This indicates that the stimuli that remained most stable across illuminants were also classified most similarly across the different observers.

Our data do not address the causal relationship between observer consistency and color constancy, but the fact that the correlation remained significant even when taking category size effects into account indicates that this pattern is not only caused by the physical interactions between illuminant and surface reflectance spectra.

Effect of local contrast

Changing the background from gray to black and making observers wear black gloves had virtually no effect on color classification consistency for real surfaces. We devised this manipulation to resemble the manipulation in our previous experiment with monitor simulations, where constancy dropped when the monitor

background was set to black instead of setting it to the chromaticity of the illuminant. Because of the general differences between these two studies—the simulated stimuli were presented on a monitor in a viewing booth, while the stimuli here were presented on the table in a regular room—the cue manipulations necessarily had a different impact on the general viewing context. In the former experiment, the background manipulation introduced a clear discontinuity between the monitor and the illuminated wall. In the present experiment, the local context change introduced by the black gloves and the black cloth was not as dramatic because the scene was richer to begin with. Kraft et al. (2002) found with real displays that scene complexity did not affect constancy when there were many valid cues to the illuminant in the scene, but when the number of cues decreased, increasing scene complexity did improve performance. In addition, Brainard (1998) and Kuriki and Uchikawa (1998) showed that a local contrast manipulation did not have a large effect on constancy in a real scene. It appears that our real scene was rich enough that reducing cues from local contrast was not enough to hamper performance.

Color classification as a method for measuring color constancy

Color constancy has been traditionally measured with asymmetric matching, where stimuli are matched in color across two different illumination contexts (e.g., Arend & Reeves, 1986; Bäuml, 1999; Brainard et al., 1997), and with achromatic matches where a stimulus is made to look achromatic under different illuminants (e.g., Brainard, 1998; Helson & Michels, 1948; Werner & Walraven, 1982). Recently, new tasks to measure constancy have been introduced to overcome some of the caveats of matching methods (e.g., Bramwell & Hurlbert, 1996; Craven & Foster, 1992; Khang & Zaidi, 2002). These operational (Craven & Foster, 1992) or forced-choice (Bramwell & Hurlbert, 1996; Khang & Zaidi, 2002) tasks appear to be more directly related to the functional role of color constancy in helping to identify objects across changes in viewing context. The forced-choice method advocated by Zaidi et al. (Khang & Zaidi, 2002; Robilotto & Zaidi, 2004; Zaidi & Bostic, 2008) was devised to reflect the fact that even though the appearance of object colors may change from one illuminant context to the next, perceived surface reflectance does not necessarily have to change. A similar principle is reflected in the differentiation of a "surface match" task and an "appearance match" task in asymmetric matching experiments, where some groups have found constancy to be significantly higher for the surface matches than for the appearance matches (e.g., Arend & Reeves, 1986; Kuriki & Uchikawa, 1996; Troost & de Weert, 1991).

Even though there is no direct relationship between color classification and object identification, classification is arguably more similar to the tasks of everyday color vision than matching. Importantly, categorical color perception is necessarily more robust to changes in the proximal stimulus than color discrimination (see Olkkonen et al., 2009, for further discussion). This also means, however, that there will always be some baseline consistency in color categories even without chromatic adaptation. Our lower bound prediction was meant to take this baseline consistency into account. Indeed, on average, half of the stimulus collection remained in the same category under illuminant changes without assuming color constancy. Color classification performance across illuminants was much closer to test–retest reliability than to the lower bound prediction, however, indicating that categorical color constancy was close to its upper bound for our experimental conditions.

Color naming consistency across illuminant changes has been previously measured for simulated (Troost & de Weert, 1991) and for unsaturated real (Granzier et al., 2009) papers. Troost and de Weert (1991) reported that on average 38% of the simulated papers were classified in the same category over five illuminants. Granzier et al. (2009) reported an average correct identification rate of 55% for their six samples across 12 viewing contexts. In the present study, the proportion of chips classified in the same category under all five illuminants ranged from 43% to 60%; naming consistency (similar to the identification rate in Granzier et al.'s study), ranged from 78% to 86%. The different degrees of consistency found by Troost and de Weert, on the one hand, and by Granzier et al. and by us, on the other hand, probably reflect a difference between real and simulated stimuli. The fact that we found higher consistency than Granzier et al., on the other hand, might be due to the subtle differences in quantifying consistency and to the different number of illuminant conditions employed. It appears, however, that the broad features of the data by Troost and deWeert and by Granzier et al. generalize to a larger set of stimulus papers and to a different set of illuminants.

Perhaps the largest advantage in color classification compared to other widely used methods is its speed, which makes measuring constancy for a large portion of color space feasible. This, in turn, allows detailed analysis of color constancy for each color category, as well as constancy as a function of stimulus hue, saturation, and lightness. The present results together with our previous work on simulated surfaces show that categorical color constancy is not homogeneous across the whole color space, and that most, but not all of this inhomogeneity can be accounted for by category size effects. In addition, categorical color constancy was as good for the chromatic as for the achromatic categories, with the exception of the black surfaces for which consistency was 100%. This is in agreement with Speigle and Brainard (1999), who showed that asymmetric matches for chromatic stimuli could be predicted from achromatic settings as long as the viewing conditions were held constant. Our data do point to a

slight advantage for category prototypes in terms of classification consistency, but this issue needs to be investigated further with an experiment focused on category prototypes.

Is constancy better for real stimuli?

In the few direct comparisons between real and simulated displays, conclusive differences between the two display media have not been found (Berns & Gorzynski, 1991; Brainard & Ishigami, 2005; Savoy & O'Shea, 1993). On the other hand, it appears based on indirect comparisons that constancy might be better for real scenes (Smithson, 2005). This is partly confirmed by our data: classification consistency is overall slightly higher for real surfaces than what we found previously for simulated surfaces. More importantly, however, both sets of data warrant the same conclusions: categorical color constancy is robust across illuminant changes and is comparable to within-observer and across-observer reliability.

Acknowledgments

We are grateful to Claudia Kubicek for data collection. We would also like to thank Robert O'Shea for sharing his unpublished manuscript and David H. Brainard for insightful discussions. This work was supported by the German Research Foundation Grant Ge 879/5-3.

Commercial relationships: none.
Corresponding author: Maria Olkkonen.
Email: mariaol@sas.upenn.edu.
Address: Department of Psychology, University of Pennsylvania, 3401 Walnut St., Philadelphia, PA 19104, USA.

References

Amano, K., & Foster, D. (2008). Categorical color perception in natural scenes under different illuminants [Abstract]. *Journal of Vision, 8*(6):572, 572a, http://www.journalofvision.org/content/8/6/572, doi:10.1167/8.6.572.

Arend, L. E., & Reeves, A. (1986). Simultaneous color constancy. *Journal of the Optical Society of America A, 3,* 1743–1751.

Arend, L. E., Reeves, A., Schirillo, J., & Goldstein, R. (1991). Simultaneous color constancy: Paper with diverse Munsell values. *Journal of the Optical Society of America A, 8,* 661–672.

Bäuml, K.-H. (1994). Color appearance: Effects of illuminant change under different surface collections. *Journal of the Optical Society of America A, 11,* 531–542.

Bäuml, K.-H. (1999). Simultaneous color constancy: How surface color perception varies with the illuminant. *Vision Research, 39,* 1531–1550.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley, CA: University of California Press.

Berns, R. S., & Gorzynski, M. E. (1991). *Simulating surface colors on CRT displays: The importance of cognitive clues*. Paper presented at the AIC Conference: Colour and Light, Sydney, Australia, 25–28 June.

Bloj, M. G., Kersten, D., & Hurlbert, A. C. (1999). Perception of three-dimensional shape influences colour perception through mutual illumination. *Nature, 402,* 877–879.

Boyaci, H., Doerschner, K., Snyder, J. L., & Maloney, L. T. (2006). Surface color perception in three-dimensional scenes. *Visual Neuroscience, 23,* 311–321.

Brainard, D. H. (1998). Color constancy in the nearly natural image: II. Achromatic loci. *Journal of the Optical Society of America A, 17,* 307–325.

Brainard, D. H., Brunt, W. A., & Speigle, J. M. (1997). Color constancy in the nearly natural image: I. Asymmetric matches. *Journal of the Optical Society of America A, 14,* 2091–2110.

Brainard, D. H., & Ishigami, K. (2005). Factors influencing the appearance of CRT colors. In *Proceedings of the IS&T/SID Color Imaging Conference: Color Science, Systems, and Applications* (pp. 62–66). Springfield, Va.: Society for Imaging Science and Technology.

Bramwell, D. I., & Hurlbert, A. C. (1996). Measurements of colour constancy by using a forced-choice matching technique. *Perception, 25,* 229–241.

Chichilnisky, E. J., & Wandell, B. A. (1999). Trichromatic opponent color classification. *Vision Research, 39,* 3444–3458.

Craven, B. J., & Foster, D. H. (1992). An operational approach to colour constancy. *Vision Research, 32,* 1359–1366.

de Almeida, V. M., Fiadeiro, P. T., & Nascimento, S. M. (2004). Color constancy by asymmetric color matching with real objects in three-dimensional scenes. *Visual Neuroscience, 21,* 341–345.

Delahunt, P. B., & Brainard, D. H. (2004). Does human color constancy incorporate the statistical regularity of natural daylight? *Journal of Vision, 4*(2):1, 57–81, http://www.journalofvision.org/content/4/2/1, doi:10.1167/4.2.1. [PubMed] [Article]

Foster, D. H., Amano, K., & Nascimento, S. M. C. (2006). Color constancy in natural scenes explained by global image statistics. *Visual Neuroscience, 23,* 341–349.

Granzier, J. J. M., Brenner, E., & Smeets, J. B. J. (2009). Reliable identification by color under natural conditions. *Journal of Vision, 9*(1):39, 1–8, http://www.journalofvision.org/content/9/1/39, doi:10.1167/9.1.39. [PubMed] [Article]

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]

Hedrich, M., Bloj, M., & Ruppertsberg, A. I. (2009). Color constancy improves for real 3D objects. *Journal of Vision, 9*(4):16, 1–16, http://www.journalofvision.org/content/9/4/16, doi:10.1167/9.4.16. [PubMed] [Article]

Helson, H., & Michels, W. C. (1948). The effect of chromatic adaptation on achromaticity. *Journal of the Optical Society of America, 38,* 1025–1032.

Hoffman, D. M., Girshick, A. R., Akeley, K., & Banks, M. S. (2008). Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision, 8*(3):33, 1–30, http://www.journalofvision.org/content/8/3/33, doi:10.1167/8.3.33. [PubMed] [Article]

Jameson, D., & Hurvich, L. M. (1989). Essay concerning color constancy. *Annual Review of Psychology, 40,* 1–22.

Kay, P., & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences of the United States of America, 100,* 9085–9089.

Khang, B.-G., & Zaidi, Q. (2002). Cues and strategies for color constancy: Perceptual scission, image junctions and transformational color matching. *Vision Research, 42,* 211–226.

Kraft, J. M., & Brainard, D. H. (1999). Mechanisms of color constancy under nearly natural viewing. *Proceedings of the National Academy of Sciences of the United States of America, 96,* 307–312.

Kraft, J. M., Maloney, S. I., & Brainard, D. H. (2002). Surface-illuminant ambiguity and color constancy: Effects of scene complexity and depth cues. *Perception, 31,* 247–263.

Kuriki, I., & Uchikawa, K. (1996). Limitations of surface-color and apparent-color constancy. *Journal of the Optical Society of America A, 13,* 1622–1636.

Kuriki, I., & Uchikawa, K. (1998). Adaptive shift of visual sensitivity balance under ambient illuminant change. *Journal of the Optical Society of America A, 15,* 2263–2274.

Ling, Y., Allen-Clarke, L., Vurro, M., & Hurlbert, A. C. (2008). The effect of object familiarity and changing illumination on colour categorization. *Perception, 37,* 149.

Ling, Y., & Hurlbert, A. (2008). Role of color memory in successive color constancy. *Journal of the Optical Society of America A, 25,* 1215–1226.

Linhares, J. M., Pinto, P. D., & Nascimento, S. M. (2008). The number of discernible colors in natural scenes. *Journal of the Optical Society of America A, 25,* 2918–2924.

Lucassen, M. P., & Walraven, J. (1996). Color constancy under natural and artificial illumination. *Vision Research, 36,* 2699–2711.

MacLaury, R. E. (1997). Ethnographic evidence of unique hues and elemental colors. *Behavioral and Brain Sciences, 20,* 202–203.

Marín-Franch, I., & Foster, D. H. (2010). Number of perceptually distinct surface colors in natural scenes. *Journal of Vision, 10*(9):9, 1–7, http://www.journalofvision.org/content/10/9/9, doi:10.1167/10.9.9. [PubMed] [Article]

Murray, I. J., Daugirdiene, A., Vaitkevicius, H., Kulikowski, J. J., & Stanikunas, R. (2006). Almost complete colour constancy achieved with full-field adaptation. *Vision Research, 46,* 3067–3078.

Nickerson, D., & Newhall, S. M. (1943). A psychological color solid. *Journal of the Optical Society of America, 33,* 419–422.

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]

Philipona, D. L., & O'Regan, J. K. (2006). Color naming, unique hues, and hue cancellation predicted from singularities in reflection properties. *Visual Neuroscience, 23,* 331–339.

Pointer, M. R., & Attridge, G. (1998). The number of discernible colors. *Color Research and Application, 23,* 52–54.

Reeves, A. J., Amano, K., & Foster, D. H. (2008). Color constancy: Phenomenal or projective? *Perception & Psychophysics, 70,* 219–228.

Regier, T., Kay, P., & Cook, R. S. (2005). *Universal foci and varying boundaries in linguistic color categories.* Paper presented at the 27th Meeting of the Cognitive Science Society, Stresa, Italy.

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.

Rinner, O., & Gegenfurtner, K. R. (2000). Time course of chromatic adaptation for color appearance and discrimination. *Vision Research, 40,* 1813–1826.

Robilotto, R., & Zaidi, Q. (2004). Limits of lightness identification for real objects under natural viewing conditions. *Journal of Vision, 4*(9):9, 779–797, http://www.journalofvision.org/content/4/9/9, doi:10.1167/4.9.9. [PubMed] [Article]

Savoy, R. L., & O'Shea, R. P. (1993). Color constancy with reflected and emitted light. *Perception, 22,* 61.

Schultz, S., Doerschner, K., & Maloney, L. T. (2006). Color constancy and hue scaling. *Journal of Vision, 6*(10):10, 1102–1116, http://www.journalofvision.org/content/6/10/10, doi:10.1167/6.10.10. [PubMed] [Article]

Smithson, H. (2005). Sensory, computational and cognitive components of human color constancy. *Philosophical Transactions of the Royal Society of London B, 360,* 1329–1346.

Smithson, H., & Zaidi, Q. (2004). Colour constancy in context: Roles for local adaptation and levels of reference. *Journal of Vision, 4*(9):3, 693–710, http://www.journalofvision.org/content/4/9/3, doi:10.1167/4.9.3. [PubMed] [Article]

Snyder, J. L., Doerschner, K., & Maloney, L. T. (2005). Illumination estimation in three-dimensional scenes with and without specular cues. *Journal of Vision, 5*(10):8, 863–877, http://www.journalofvision.org/content/5/10/8, doi:10.1167/5.10.8. [PubMed] [Article]

Speigle, J. M., & Brainard, D. H. (1996). Is color constancy task independent? In *Proceedings of the 4th IS&T/SID Color Imaging Conference, Scottsdale, AZ* (pp. 167–172).

Speigle, J. M., & Brainard, D. H. (1999). Predicting color from gray: The relationship between achromatic adjustment and asymmetric matching. *Journal of the Optical Society of America A, 16,* 2370–2376.

Troost, J. M., & de Weert, C. M. (1991). Naming versus matching in color constancy. *Perception & Psychophysics, 50,* 591–602.

Uchikawa, K., Uchikawa, H., & Boynton, R. M. (1989). Partial color constancy of isolated surface colors examined by a color-naming method. *Perception, 18,* 83–91.

Webster, M. A., & Kay, P. (2005). Variations in color naming within and across populations [commentary on Steels and Belpaeme]. *Behavioral and Brain Sciences, 28,* 512–513.

Webster, M. A., & Kay, P. (2007). Individual and population differences in focal colors. In R. E. MacLaury, G. V. Paramei, & D. Dedrick (Eds.), *The anthropology of colors* (pp. 29–53). Amsterdam, The Netherlands: John Benjamins.

Werner, A. (2006). The influence of depth segmentation on colour constancy. *Perception, 35,* 1171–1184.

Werner, J. S., & Walraven, J. (1982). Effect of chromatic adaptation on the achromatic locus: The role of contrast, luminance and background color. *Vision Research, 22,* 929–943.

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

Yang, J. N., & Shevell, S. K. (2002). Stereo disparity improves color constancy. *Vision Research, 42,* 1979–1989.

Yang, J. N., & Shevell, S. K. (2003). Surface color perception under two illuminants: The second illuminant reduces color constancy. *Journal of Vision, 3*(5):4, 369–379, http://www.journalofvision.org/content/3/5/4, doi:10.1167/3.5.4. [PubMed] [Article]

Zaidi, Q., & Bostic, M. (2008). Color strategies for object identification. *Vision Research, 48,* 2673–2681.