Chromatic settings and the structural color constancy index

Jordi Roca-Vila   Computer Vision Center, Bellaterra, Barcelona, Spain

C. Alejandro Parraga   Computer Vision Center/Computer Science Department, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

Maria Vanrell   Computer Vision Center/Computer Science Department, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

Color constancy is usually measured by achromatic setting, asymmetric matching, or color naming paradigms, whose results are interpreted in terms of indexes and models that arguably do not capture the full complexity of the phenomenon. Here we propose a new paradigm, *chromatic setting*, which allows a more comprehensive characterization of color constancy through the measurement of multiple points in color space under immersive adaptation. We demonstrated its feasibility by assessing the consistency of subjects' responses over time. The paradigm was applied to two-dimensional (2-D) Mondrian stimuli under three different illuminants, and the results were used to fit a set of linear color constancy models. The use of multiple colors improved the precision of more complex linear models compared to the popular diagonal model computed from gray. Our results show that a diagonal plus translation matrix that models mechanisms other than cone gain might be best suited to explain the phenomenon. Additionally, we calculated a number of color constancy indices for several points in color space, and our results suggest that interrelations among colors are not as uniform as previously believed. To account for this variability, we developed a new structural color constancy index that takes into account the magnitude and orientation of the chromatic shift in addition to the interrelations among colors and memory effects.

Introduction

Color constancy is a perceptual phenomenon that keeps the color of objects relatively stable under varying illumination conditions (Foster, 2011; Land, 1964; Smithson, 2005). Since a full recovery of the spectral properties of either the illumination or objects by the trichromatic eye does not have a unique solution, several possible strategies have been proposed to make color constancy possible. These include restrictions on the number and dimensionality of the spectral reflectances and illuminants available (Maloney & Wandell, 1986), normalizations with respect to the illumination (Brainard & Wandell, 1986), assumptions about the brightest visible object (Land & McCann, 1971) or the average color of the world (Buchsbaum, 1980), higher order statistical properties of the environment and other regularities (Golz & MacLeod, 2002; Hordley, 2006), or a combination of these. However, none of the explanations proposed so far provides a complete representation of how a visual scene is perceived under an illumination shift in naturalistic, complex, unconstrained conditions. For instance, the degree of color constancy may depend on internal criteria derived from different judgments of the scene, as demonstrated by the hue saturation versus paper matches of Arend and Reeves (1986). Other confounds may depend on the ability of subjects to attribute changes in the scene to either changes in the spectral composition of the illuminant or the reflecting properties of objects in that scene (Foster & Nascimento, 1994). High level visual memory may also play an important role in judgments of surface color, as demonstrated by Hansen, Olkkonen, Walter, and Gegenfurtner (2006).

The degree and quality of the color constancy experienced by observers is usually measured by a variety of psychophysical techniques. A typical experiment compares the colors an observer perceives under a given state of illuminant adaptation to the colors perceived under another state, and the differences are then interpreted using models and indices (Brainard,

Brunt, & Speigle, 1997; Foster, 2011). Models attempt to predict the color appearance of other nonmeasured colors while indices quantify the degree of color constancy achieved. The most popular color constancy paradigms are achromatic setting, asymmetric color matching, and color naming (Foster, 2011; Smithson, 2005). Achromatic setting measures the perceptual stability of the achromatic locus under a change of adaptation by asking subjects to modify a stimulus until it appears "achromatic." It has been pointed out that this is a local measurement that may or may not be influenced by manipulations of other regions of the scene and also that one measure may not be enough to estimate the stability of perceived colors away from the neutral point (Delahunt & Brainard, 2004; Foster, 2003, 2011; Schultz, Doerschner, & Maloney, 2006). Asymmetric color matching (Arend & Reeves, 1986; Wyszecki & Stiles, 1982) compares binocular or dichoptical stimuli under different illuminants, presented either simultaneously or successively. Subjects adjust a patch under one illumination to match another under a different illumination. This method requires that the state of adaptation follows closely the change of illumination, a strong assumption especially in the case of alternate viewing paradigms (Foster, 2011). Color naming paradigms rely on the subjects' internal color categories by asking them to classify samples under different illuminants. It has been argued that color naming provides a more direct method for measuring color constancy (Foster, 2011) on the grounds that it is less sensitive to the instructions given to subjects (Arend & Reeves, 1986; Troost & de Weert, 1991). The main setback of the last method is the large number of discernible colors (more than two million), much larger than the number of possible names (Linhares, Pinto, & Nascimento, 2008; Pointer & Attridge, 1998), resulting in limited accuracy (Foster, 2011). Variants include determining unique hues and estimating the degree of color constancy from the response categories of large numbers of samples and the position of color boundaries (Chichilnisky & Wandell, 1999; Smithson & Zaidi, 2004) under different states of adaptation (Hansen, Walter, & Gegenfurtner, 2007; Kulikowski & Vaitkevicius, 1997; Olkkonen, Hansen, & Gegenfurtner, 2009; Olkkonen, Witzel, Hansen, & Gegenfurtner, 2010).

Color constancy from multiple points in color space

In addition to using internal gray as a reference, some researchers have included multiple color references to study color constancy (Hansen et al., 2007; Kulikowski & Vaitkevicius, 1997; Olkkonen et al., 2009; Olkkonen et al., 2010) and to determine

properties such as the boundaries between color categories (Benavente, Parraga, & Vanrell, 2009; Smithson & Zaidi, 2004). Some of these studies have measured directly the color appearance of several colored patches under different illuminants (Kulikowski & Vaitkevicius, 1997; Speigle & Brainard, 1997), while others have used color naming to derive a conclusion about the categorical structure of color space (Hansen et al., 2007; Olkkonen et al., 2009; Olkkonen et al., 2010; Troost & de Weert, 1991). In the direct measures, immediate color constancy seems to hold best for hues corresponding to "typical" colors as compared with the adjacent hues; however, this effect may be residual (Kulikowski & Vaitkevicius, 1997). Through the use of color naming techniques and a large set of colored samples, Hansen et al. (2007) and Olkkonen and colleagues (Olkkonen et al., 2009; Olkkonen et al., 2010) achieved different levels of color constancy according to the degree of information provided. They modeled the transformations of the perceptual color space under different illuminations by computing the boundaries of the color categories (Hansen et al., 2007; Olkkonen et al., 2009) and computing the color constancy indices of the categorical prototypes (Olkkonen et al., 2010). Their conclusions were that the categorical structure of color space has a high degree of robustness under changes of illumination, which could be explained by linear models. However, Hansen et al. (2007) reported small rotations away from the illumination color.

In this work we explore whether keeping adaptation constant throughout a single session and characterizing colors other than gray improves the accuracy of color constancy modeling and if that is so, how to use this knowledge to further our understanding of the phenomenon. With this in mind, we have developed a color constancy paradigm that tries to minimize the weaknesses while keeping the strongest points of previous paradigms. In our method, the measurements are done under a permanent state of adaptation, thus avoiding potential illuminant-switching issues. Our paradigm can be seen as an extension of the achromatic setting paradigm, which, instead of using only the internal "gray" reference, uses several categorical colors, exploiting the ability of subjects to consistently replicate focal colors over time. Focal colors (Berlin & Kay, 1969/1991; Boynton & Olson, 1987) are by definition the most representative colors of each naming category, and there is strong evidence of the effect of language (Heider, 1972; Kay et al., 2009) and memory (Hansen et al., 2006; Ling & Hurlbert, 2008; Nemes, McKeefry, & Parry, 2010) on their perceptual categorization. Although the ability of subjects to match a memorized color decreases in general with increasing interstimulus intervals (Nemes et al., 2010), there is some evidence that focal colors can be

remembered more accurately than other colors (Heider, 1972). Other effects, such as a tendency to remember more saturated or brighter/dimmer colors have been linked to color constancy (Jin & Shevell, 1996; Ling & Hurlbert, 2008). Since our paradigm relies strongly on color memory ability, we studied its validity by testing the stability of these internal references along the experiment. To test whether our paradigm provides a more comprehensive measure of the color constancy phenomenon, we applied linear models to study its behavior, e.g., to what extent these models are capable of absorbing the growing data complexity that results from the addition of extra measurements. Finally, we developed a new color constancy index that arguably captures the intrinsic complexity of the phenomenon in a single value, while still in agreement with the previous color constancy literature.

Methods

Overview

We present a new psychophysical paradigm to determine color constancy under immersive illumination that measures the perception of nine colors under different states of adaptation by using CRT monitor based stimuli. The stimuli were combinations of three different 2-D Mondrians and three different illuminants. The subject's task was to select and then reproduce a particular color from memory. In order to rule out memory failings the experimental procedure included a series of repeatability tests.

The chromatic setting paradigm

Our paradigm consists of two steps as illustrated by Figure 1. In the first step, subjects were asked to select colors that best represented basic color terms within a limited region of the color space (*Bounding Cylinder*, represented by a red circle). These were gray, green, blue, purple, pink, red, brown, orange, and yellow (Berlin & Kay, 1991). The squares within the red circle in Figure 1 symbolize the colors selected during this first step, which we called *reference session*. We termed these colors *Selected Representatives* (SRs). In the second step, which we called *regular session,* the same subjects were asked to reproduce these SRs under different conditions of background and illumination. The squares outside the red circle in Figure 1 correspond to these colors, and the arrow represents the change in adaptation state. Since the new paradigm can be seen as an extension of the achromatic setting
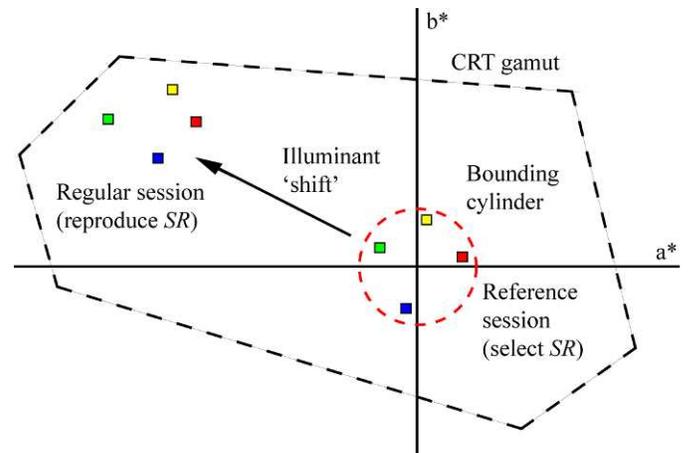
Figure 1. Schematics of the chromatic setting paradigm in the a*b* plane of CIELab color space. The black broken lines represent the boundary of the CRT gamut. The squares inside the red circle represent the colors selected in the reference session. The squares outside the red circle represent the colors reproduced once adapted to the new illuminant in a regular session. The arrow represents the chromatic shift induced by the illumination.

paradigm to multiple colors, we named it *Chromatic Setting.*

The red circle in Figure 1 corresponds to the projection of a cylinder in the a*b* plane. This cylinder was introduced to limit subjects choices, thus avoiding highly saturated colors that fall outside the CRT gamut when "illuminated." Details on the Bounding Cylinder implementation can be found below.

Experimental setup

All sessions were conducted inside a dark room, with all walls lined in black. The experiment was programmed in Matlab and the stimuli were displayed on a CRT Mitsubishi Diamond Pro 2045SU monitor (Mitsubishi Group, Tokyo, Japan) at 100Hz, driven by a ViSaGe graphics card (Cambridge Research Systems Ltd., Rochester, UK) with 12 bits color resolution per channel. The CRT screen measured 389 mm in height by 292 mm in width subtending approximately $22° \times 17°$ and was the only light source in the room. Its resolution was 1024 $\times$ 768 pixels. Viewing was binocular and unrestrained. The monitor was calibrated regularly using a Minolta *ColorCal* colorimeter (Konica Minolta, Tokyo, Japan) and CRS software. We used the COLORLAB (Malo & Luque, 2002) toolbox to get the color space conversions needed. Subjects modified the test stimuli by navigating the CIELab color space using six different buttons, two for each color space dimension on a commercial gamepad. The reference white point was D65, luminance = 100 cd/m^2.

Subjects

Ten subjects, six male and four female, took part in our experiments. They were between 20 and 44 years old, and their color vision was normal as tested by the Ishihara color vision test (Ishihara, 1972) and the Farnsworth-Munsell D15 hue test (Farnsworth, 1957). All had self-reported normal or corrected-to-normal visual acuity. Three of the subjects were the authors. The rest were naive to the purposes of the experiment, and of these, three were paid.

Stimuli

Our basic stimulus consisted of a Mondrian background pattern, i.e., a set of randomly overlaid colored rectangles, distributed across the screen. The average rectangle size was 50×50 pixels. There were three types of backgrounds:

Type 0

It was built from seven intensity levels of the same D65 chromaticity. They were equally spaced between 40 and 70 Lab lightness units and their luminances in cd/m^2 were: 11.25, 14.54, 18.42, 22.93, 28.12, 34.05 and 40.75. Its mean was 22.66 cd/m^2.

Type I

It was built from the SRs chosen by each subject in reference sessions (see details below). There were eight colors in total: green, blue, purple, pink, red, brown, orange, and yellow. Their averaged luminance range was between 12.77 and 39.29 cd/m^2, mean = 25.11 cd/m^2.

Type II

It was built from eight hues halfway between those of Type I, with similar saturation and lightness: blue-purple, purple-pink, purple-red, red-orange, orange-yellow, orange-brown, yellow-green, and green-blue. Their averaged luminance range was between 16.87 and 35.54 cd/m^2, mean = 24.35 cd/m^2.

The number and sizes of rectangles were manipulated so that the pixel average chromaticity of all background types prior to illumination was that of D65. Background Types I and II did not contain achromatic D65 rectangles to avoid giving the observer cues about the illuminant (Foster, 2011). Unique randomized Mondrians were created for each experimental trial: No observer saw the same Mondrian twice. To illuminate the Mondrian pattern, we first assigned to each rectangle a spectral reflectance function with the same XYZ tristimulus values as the desired colors. These reflectance functions were obtained by interpolating linearly be-

Illuminant	x	y
D65	0.312	0.329
Greenish	0.296	0.453
Yellowish	0.453	0.434

Table 1. CIE *xy* chromaticity of the illuminants.

tween real spectral reflectances from a large set of Munsell chips assuming a Lambertian reflectance model—see COLORLAB (Malo & Luque, 2002) for implementation details. Illumination was simulated by performing the spectral product of each rectangle's reflectance by one of three illuminants (*D65*, *greenish* and *yellowish*), whose CIE *xy* chromaticities are shown in Table 1. The luminance range in cd/m^2 for the illuminated stimuli was between 11.25 and 40.74 for the D65 illuminant; between 11.24 and 40.73 for the greenish illuminant, and between 11.20 and 40.56 for the yellowish illuminant. The mean values in cd/m^2 were 24.04, 23.7, and 24.37, respectively.

Procedure

The experiment consisted of sixteen sessions divided in three groups: *reference*, *regular*, and *repeatability tests*. Figure 2 shows the time sequence of the experiment. First there was a *training* period followed by the *reference session*, after which the main body of the experiment started. It consisted of nine *regular sessions* and three interleaved repeatability tests (occurring at the beginning, halfway, and at the end of the regular sessions) whose aim was to track variations in subject's responses. Subjects completed all experiments in less than three weeks, and no more than two sessions per day were allowed. Details of the different sessions were as follows:

Reference session

It consisted of a single session with Type 0 background and D65 illumination, and it started just after the training was completed. Subjects were instructed to select the most representative colors for each of the eight basic chromatic categories. The choice of available colors was constrained by the Bounding Cylinder (see squares within the red circle in Figure 1).

Regular sessions

They consisted of nine sessions combining the three illuminants and three background types described before. Each regular session followed a similar protocol as the reference session, except that subjects were instructed to reproduce the same SRs they had selected
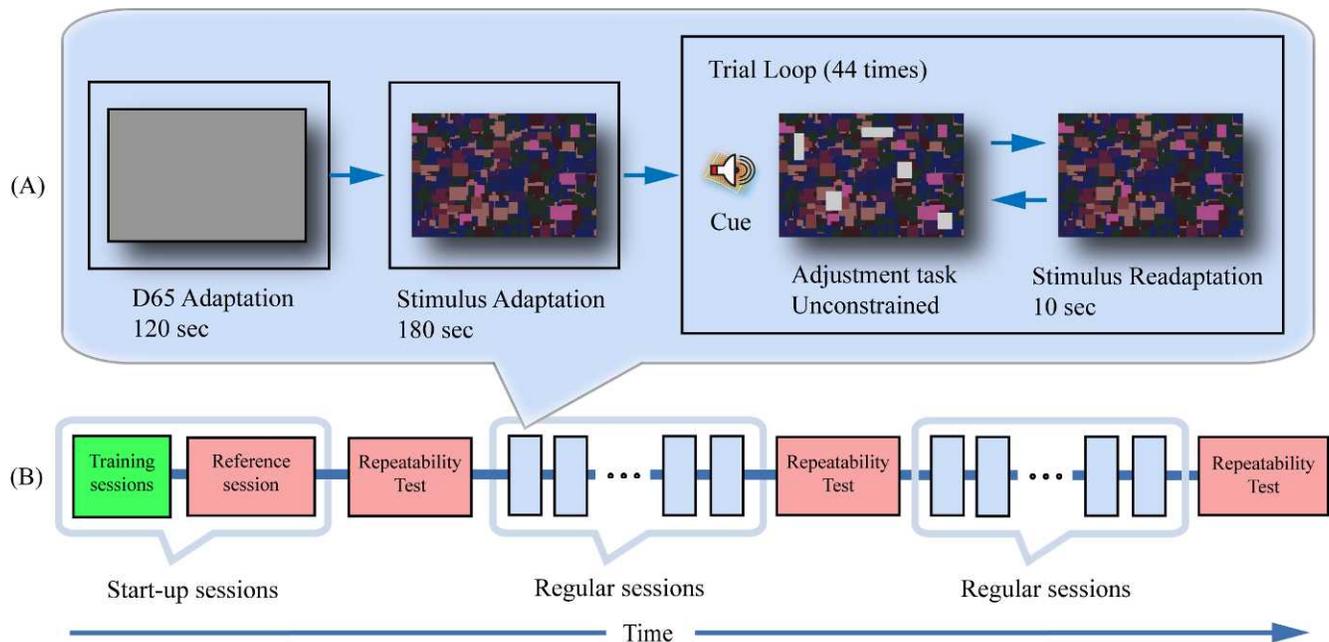
Figure 2. Temporal sequence of the experiment. Panel A shows the common schematics for a reference or regular session. Panel B illustrates the setup of the whole experiment. Start-up sessions consisted of both training and reference sessions. In a reference session, subjects selected their most representative color for each category. Regular sessions were similar, except that subjects had to reproduce the same colors they had chosen in the reference session. Repeatability tests were designed to assess subject ability to reproduce the colors selected in the reference session.

in the reference session without any constraints (no Bounding Cylinder).

Repeatability test

It consisted of three groups of two sessions each. In the first session, subjects were asked to reproduce the SR chosen before, this time under Type 0 background, D65 illuminant, and within the Bounding Cylinder. This is equivalent to a reference session where subjects reproduce instead of selecting the colors. The second session was a regular session with Type II background and greenish illumination. This choice was arbitrary: We decided to include only one illumination in order to avoid extending the experiment unnecessarily, while keeping the same experimental complexity as in the regular sessions.

Training

Training occurred at the very beginning and consisted of repeating two consecutive sessions: a reference session followed by a regular session both with Type 0 background and D65 illuminant (i.e., in the second session there was no Bounding Cylinder). The objective of this was for subjects to understand the different instructions in both cases. Pilot sessions with the authors as subjects showed that in regular sessions it was possible to reach a precision of 5 ΔE^* at

reproducing the same colors after about two visits to our lab (four sessions), and this did not improve significantly afterwards. We used this value as a criterion to determine the end of training.

Panel A of Figure 2 shows the common schematics of the reference and regular experimental sessions. Each session started with a 120-s adaptation to a uniform D65 screen (luminance equal to 30 cd/m^2) followed by 180 s of adaptation to a Mondrian under the same simulated illumination to be used later in session. After that, subjects were prompted auditorily and visually (by a word written in black at the bottom of the screen) to the color category requested, and they manipulated the gamepad to either select or reproduce the colors according to their instructions. Each trial ended by pressing a "next trial" button on the gamepad which followed re-adaptation to a geometrically randomized version of the original Mondrian and illuminant for 10 s before proceeding to the next trial. There were 44 trials: In the first four, subjects were asked to produce "gray," and in the following, they were asked to produce the other eight colors five times each in random order. Test patches occurred simultaneously at multiple random locations in the Mondrian and were adjusted by the observer with no time constraints. They were spatially distributed in a random manner in every trial with the aim of forcing subjects to average test locations, thus reducing local chromatic induction effects (Otazu, Parraga, & Vanrell,
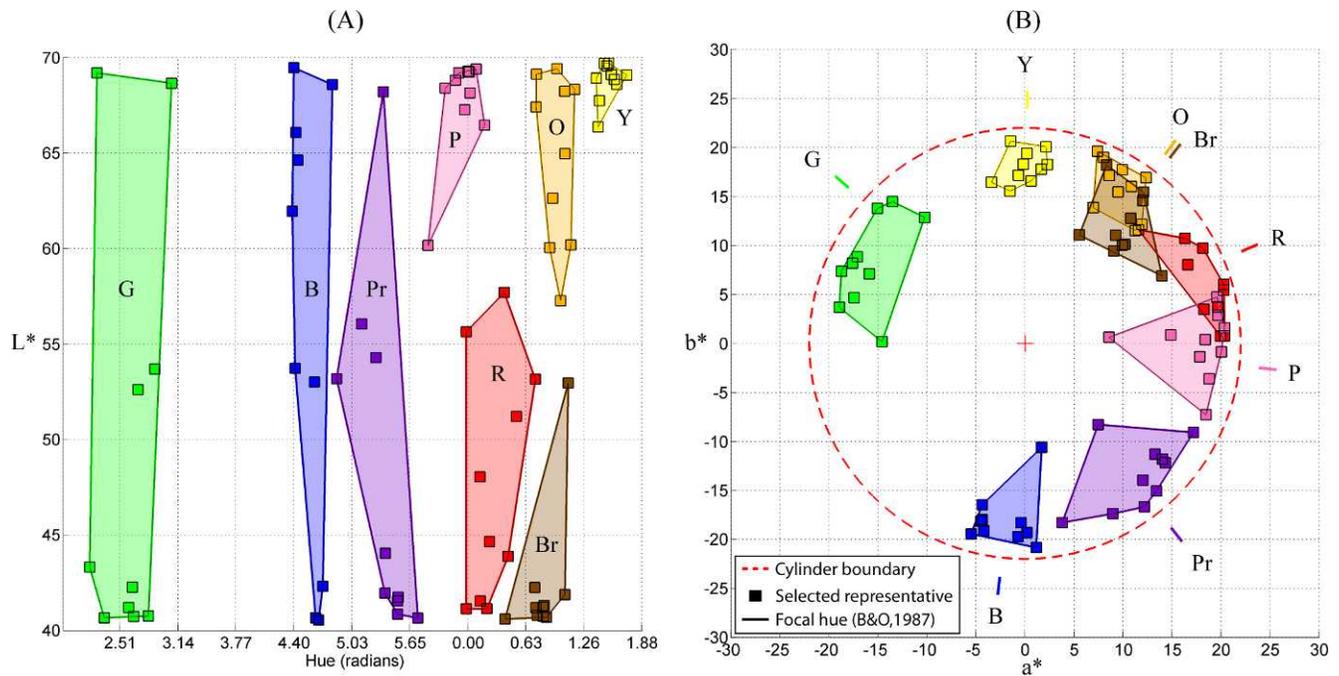
Figure 3. CIELab locations of the selected representatives adjusted in the reference sessions by all 10 subjects. Square markers in both panels indicate the average location (five trials) of each color category and subject. Color categories are labeled and color-coded with their representative colors (R-red; G-green; B-blue; Y-yellow; N-gray; W-white; K-black; P-pink; O-orange; Pr-purple; Br-Brown). Panel A shows the projection of the data in hue and lightness. Panel B shows the same data projected on the a*b* plane. The red circle shows the boundary constraints imposed by the method in the reference sessions.

2010; Shevell & Wei, 2000). The number of test patches was determined according to the following constraints: (a) the total area occupied by the test patches was between 4% and 7% of the display and (b) the pixel average chromaticity of the screen prior to illumination was equal to D65. This resulted in different number and sizes for the test patches in backgrounds Type 0 (where the pixel average was already neutral) and Type I and II backgrounds. As a consequence, the number of test patches followed a normal distribution around 25 (2.4 SD) for the Type 0 backgrounds and 4.1 (0.75 SD) for the Type I and II backgrounds.

In the cases where "gray" was requested, we randomized the chromaticity of the initial test patches around the expected value to avoid influencing the subject's response—see Brainard's *basic starting rule* (Brainard, 1998). In all other cases, the starting value of the test patches was randomly distributed around each subject's selected "gray." To obtain a single measure of a SR color, we averaged its individual trials adjustments. Each trial lasted approximately 30 s and each session approximately 25 min.

Bounding Cylinder

In the reference sessions, the palette of possible colors was limited in saturation and lightness by a cylinder whose main axis was the lightness dimension

of CIELab (L* between 30 and 70 and radius equal to 22 ΔE*). The purpose of the cylinder was strictly technical as illustrated in Figure 1: We wanted subjects to find reasonably representative samples while still allowing these colors to be "illuminated" later without exceeding the CRT-monitor gamut. This limitation and the shape of the monitor's gamut in CIELab also determined our choice of illuminants. The value of 22 ΔE* for the radius was chosen after our own (unpublished) measurements indicated that colors closer than 12 ΔE* to the achromatic locus were usually categorized as "gray." Subjects naturally tended towards choosing saturated colors, and to stop them from using the borders of the cylinder as a reference, i.e., to increase saturation until hitting the cylinder limit, the experimental program "bounced back" the stimulus inside the cylinder by a small random amount once the boundary was reached. The Bounding Cylinder was not present in regular sessions.

Results

Selected representatives and their repeatability

Figure 3 shows the CIELab location of selected representatives chosen by all subjects (D65 was used as

a reference white point). Panel A shows the data projection into the lateral surface of the Bounding Cylinder and panel B shows their projection into the a*b* plane. The limits of the Bounding Cylinder are shown as a red circle in panel B. The colored areas highlight the inter-subject variability, which is largest in the lightness dimension (Foster, 2011; Webster & Kay, 2007), particularly for green, blue and purple. From the two panels it can be inferred that there is no volumetric overlap among the different colored areas, i.e., subjects were consistent in selecting colors within categories. The figure also shows good agreement between the hue locations of our categories and the hues of Boynton and Olson's (1987) focals , plotted beyond the cylinder boundaries in panel B.

Along the experiment we kept track of the accuracy of responses over time by means of the repeatability tests as detailed in the Methods section and Figure 2B. These were conducted regularly at approximately three-day intervals and included a reference session where observers were asked to reproduce the original SR colors. Plots in Figure 4 were arranged in rows and columns. Columns correspond to two typical subjects (XO and AB) and the most inconsistent subject (LC) over time. Rows correspond to measurements taken over three-day intervals. The first row corresponds to the chromatic settings of the reference sessions and the others (rows B, C, and D) correspond to the repeatability tests in temporal sequence.

To find out whether all observers could reproduce the colors from memory within a reasonable JND range we applied a one-way ANOVA to each category across different days in each CIELab dimension and computed ΔE^* distances among the means. To obtain information about which pairs of means are significantly different, and which are not, we applied a multiple comparison procedure: the *Tukey–Kramer* method (Hochberg & Tamhane, 1987), which returns a set of pairwise comparison results. For example, to assess the repeatability of subject's XO "red" settings we considered data from the first column (rows B, C, and D) in Figure 4. These consist of three groups of five points each in the three CIELab dimensions. We applied our tests to each dimension separately, obtaining the values of $F(2, 12) = 2.25$ with $p = 0.15$ for a*, $F(2, 12) = 0.77$ with $p = 0.48$ for b*, and $F(2, 12) = 18.8$ with $p = 0.0002$ for L*. After applying the Tukey–Kramer post-hoc comparison we obtained three sets (one for each dimension) of values showing whether the "red" measures in panels B, C, and D are significantly different from each other. To assess whether "red" was well remembered we computed in all CIELab dimensions, the percentage of cases that were significantly different (in the example, observer XO could remember "red" in 78% of the cases). We repeated this procedure for all color categories, obtaining an average of 15%

significantly different measures for all subjects. There were 17% significantly different measurements for red, green, and orange, and less than 16% significantly different measurements for the other colors. The mean distance among chromatic settings within the same category was 1.79 ΔE^* for all observers and categories considered (see Table 2 below).

Some subjects complained that red and/or orange selections were not saturated enough to be called "representatives." Crucially, this did not seem to impair their capacity to remember the same color throughout the rest of the experiment even for close categories such as brown and pink.

Repeatability tests also contained a regular session with Type II background and greenish illuminant. Figure 5 shows a summary of these results. Each panel corresponds to the same observer as before (XO, AB, and LC), and each square marker corresponds to a measurement taken over three-day intervals. Notice the data shift corresponding to the change of illuminant. We applied the same approach as before and found that the means of the results populations were different in 27% of the cases and the mean distance among chromatic settings within the same category was 4.21 ΔE^* (see Table 2). This difference is likely to be due to the absence of the Bounding Cylinder, which increased uncertainty in the saturation dimension.

Although the repeatability tests show that subjects can reproduce the same SR colors, we conducted another experiment to test longer term color memory. These results, which are consistent with Figure 5, are detailed in Appendix A.

Chromatic settings under different illuminants

Figure 6 shows the averaged chromatic settings in CIELab obtained during regular sessions for all subjects, discriminated by backgrounds and separated in panels according to the illuminant. Over the regular sessions, our 10 subjects adjusted five times (four for gray) each of the nine basic colors for each of the nine different stimuli, totaling 3,960 adjustments. Only 1.4% of these adjustments were closer than five CIELab ΔE^* units from the CRT monitor gamut boundary, thus indicating that subjects did not use this boundary as a cue to find their SR colors.

As before, we plotted these results from different illuminations under the same D65 reference white point in order to highlight the amount of illumination shift, hence the displacement of the data in the plots. Figure 6 shows a tendency for subjects to choose more saturated colors in the presence of colored backgrounds than in the presence of achromatic backgrounds, i.e., squares are closer to the achromatic locus. This is true for all colors studied except for green, yellow, and orange. A
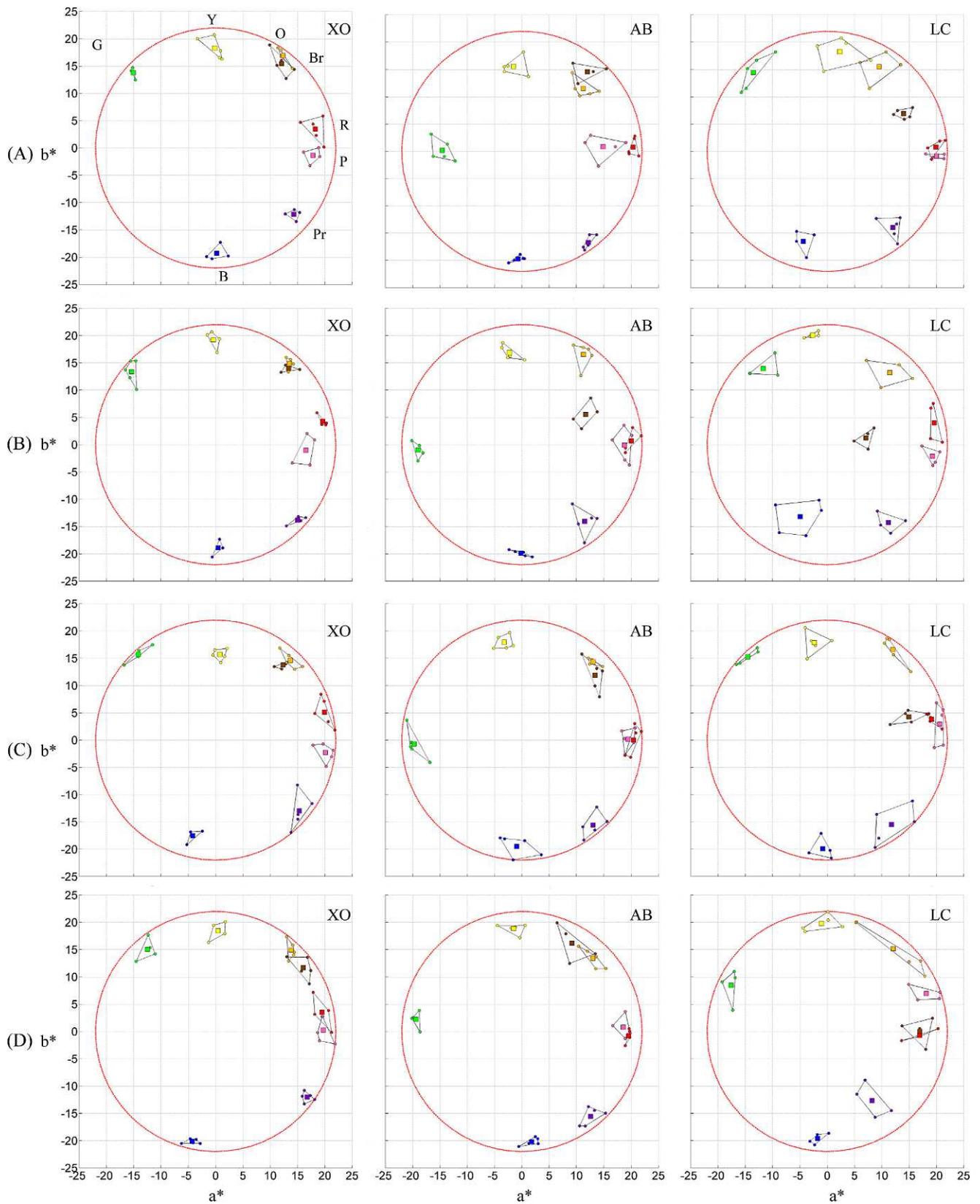
Figure 4. Chromatic settings from the reference session and the repeatability sessions. Row A shows the selected representatives chosen by three subjects in the reference session. Rows B, C, and D show the corresponding settings for the three subsequent repeatability tests. Square markers represent the average of individual trials (small dots joined by lines) and the large red circle corresponds to the Bounding Cylinder in a*b* chromaticity plane.

	Red	Green	Blue	Yellow	Neutral	Purple	Pink	Orange	Brown	Mean
First session (D65)	2.04	1.82	1.64	1.54	—	1.39	2.29	1.96	1.66	1.79
Second session (Greenish)	4.74	3.82	3.26	4.72	4.89	4.21	4.86	3.33	4.01	4.21

Table 2. ΔE^* distance among chromatic settings obtained during the repeatability test sessions. *Notes*: Values were obtained by averaging the distances between each chromatic settings and their mean. The first line corresponds to measures obtained during the first session of the repeatability test (which included a "Bounding Cylinder" and D65 illuminant), and the second line corresponds to measures obtained during the second sessions, under greenish illuminant and without a bounding cylinder. Results are discriminated by color category, and the last column shows the means.

similar outcome was reported by Brown and McLeod (1997) in their comparison between the effects of low-contrast and high-contrast multicolored surrounds. From the same figure we conclude that the type of background did not have a strong influence in the chromatic settings. However, since Types I and II backgrounds were customized for each subject according to their SR, the generalization may be masking individual effects.

Figure 7 shows a set of typical result plots, arranged in columns and rows. Each of the columns corresponds to a different illuminant and the rows to four exemplary subjects, all measured using Type II backgrounds. Inside the plots, each colored square correspond to the average of five trials (four for gray), which are shown as smaller points joined by lines. To quantify the amount of *variability* (δ) within each group of five trials we computed the average CIELab ΔE^* distance between each SR trial and the mean SR. As a white point for our calculations we used the corresponding chromaticity of each illuminant (see Table 1) at 100 cd/m^2. Since there were differences in the dispersion of data around the

mean depending on each subject and color category, we summarized δ in Table 3 where each value corresponds to the average variability over illuminant-background combinations. The average δ value was 2.09 ΔE^* (1 SD) for the reference sessions and 4.60 ΔE^* (2.06 SD) for regular sessions. The difference between these values is likely to result from the Bounding Cylinder. According to our estimations, the precision of our method is consistent with that of achromatic setting studies (Brainard, 1998), where accuracies between 4 and 5 ΔE^* are common.

The last row of Table 3 shows the mean variability δ for each color category. Some color categories such as red (mean = 4.03, 1.25 SD) and gray (mean = 4.05, 1.49 SD) have in average a smaller δ value than others, e.g., purple (mean = 5.29, 1.15 SD) and orange (mean = 5.02, 1.22 SD). These tendencies are similar across background types. However, different illuminants arguably influenced the δ value of our measures: D65 illuminant has the lowest δ value (mean = 3.83, 1.52 SD), followed by greenish (mean = 4.81, 2.08 SD), and yellowish (mean = 5.16, 2.27 SD) illuminants.
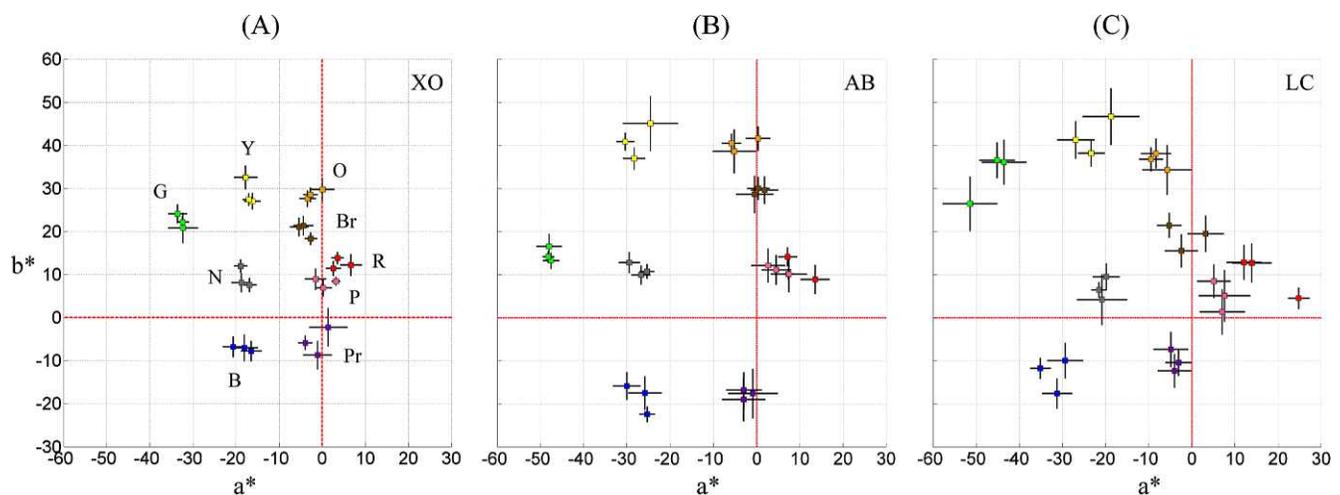
Figure 5. Chromatic settings for repeatability sessions. Results include settings for three subjects, Type II background, greenish illuminant, and no Bounding Cylinder. Each point represents the average of five trials (four for gray), and it was produced in different days over the experiment lifespan. Error bars show the standard deviation. Panels A and B correspond to typical subjects and C shows the subject with the largest variability. Notice the shift of all points towards green, corresponding to the greenish illuminant. We chose D65 as a reference white point to highlight the effects of the illuminant for illustrative purposes. Again, for clarity's sake, lightness information is not shown.
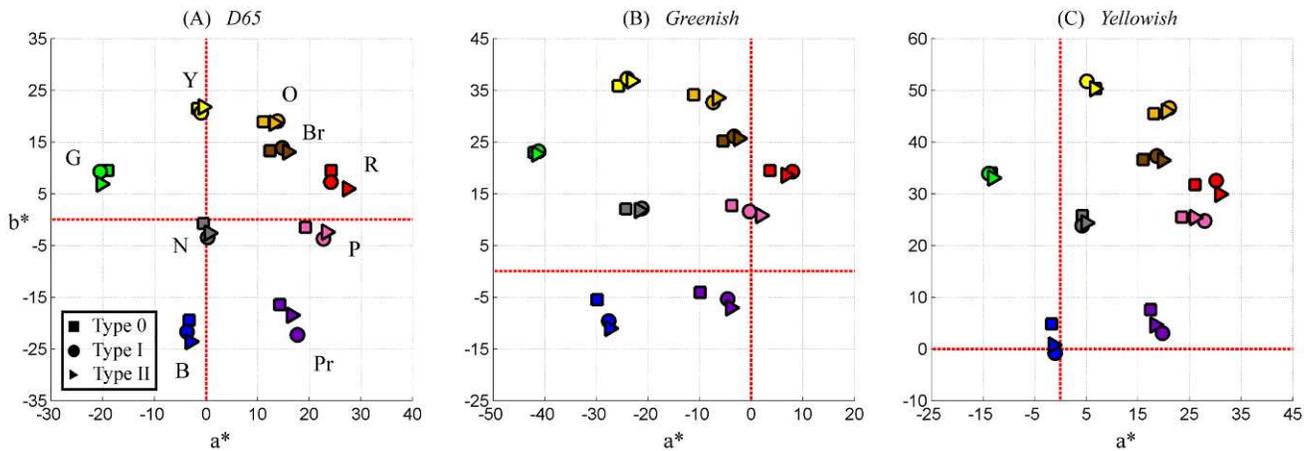
Figure 6. Average chromatic settings of the selected representatives in regular sessions. The symbols show the chromatic settings for each background type: squares for Type 0, circles for Type I, and triangles for Type II. Points were computed by averaging the corresponding SR for all subjects, for each particular background and illuminant. Panel A corresponds to D65, B to greenish, and C to yellowish illumination.

We recorded the time subjects took to complete each trial. The average was 19.5 s (5.7 *SD*) for the reference sessions and 20.7 s (6.2 *SD*) for the regular sessions. Gray took the longest to adjust (mean = 25.1, 7.6 *SD*), followed by brown (mean = 22.1, 5.6 *SD*), which in turn took longer than blue (mean = 18.6, 5.7 *SD*), purple (mean = 18.7, 4.5 *SD*), and pink (mean = 17.9, 4.7 *SD*). Red (mean = 21.1, 7.1 *SD*) and yellow (mean = 20.7, 6.2 *SD*) took longer time than pink, which was the fastest to adjust.

Color constancy indices

We quantified the extent of color constancy achieved by our subjects through three color constancy indices: the *Constancy Index* (CI; Arend, Reeves, Schirillo, & Goldstein, 1991), the *Color Constancy Index* (CCI; Ling & Hurlbert, 2008) and the *Brunswick ratio* (BR; Hansen et al., 2007; Smithson & Zaidi, 2004; Yang & Shevell, 2002), which takes into account the adaptation under the reference illumination. Equation 1 shows an example of how this was implemented for the case of BR.

When considering a particular subject's data, we noted a_c^i as the chromaticity coordinates of his/her selected representative c under illumination i (1 corresponds to D65, 2 to greenish, and 3 to yellowish). Also, b_c^i are the chromaticity coordinates of the corresponding a_c^1 when the illuminant i was applied. The numerator computes the perceptual shift, i.e., the difference between SRs chosen under D65 illuminant and greenish/yellowish illuminants. The denominator computes physical shift, i.e., the difference between SRs chosen under D65 and their chromatic coordi-

nates when illuminated by greenish/yellowish illuminants. Following this arrangement, a value of one indicates perfect color constancy and zero no color constancy.

$$\overline{\text{BR}}_c^i = \frac{||a_c^1 - a_c^i||_2}{||b_c^1 - b_c^i||_2} \quad \text{where } i = 2, 3 \text{ and}$$

$$c = 1, \dots, 9 \qquad (1)$$

Although there is no assumption of any specific color space in the index formulae, we choose CIE1976 uv, a perceptually uniform space which does not incorporate any white point normalization as CIELab does (Brainard, 1998; Wyszecki & Stiles, 1982). Table 4 shows the values of three indices (CI, $\overline{\text{BR}}$, and CCI) averaged for all subjects and considering each color category and illumination. The data highlights the discrepancies between indices, e.g., maximum and minimum values within each column (highlighted in bold) do not coincide for the same categories. In other words, the results obtained for each color category depend on the color constancy index selected.

Linear color constancy models

As Table 4 indicates, the chromatic settings of our subjects were different for different illuminants. We modeled the effects of the illuminant change using linear models of color constancy, i.e., a linear transformation matrix that relates two chromatic settings of the same color under different illuminants. To be able to relate the parameters of our models to properties of the human visual system, we chose to
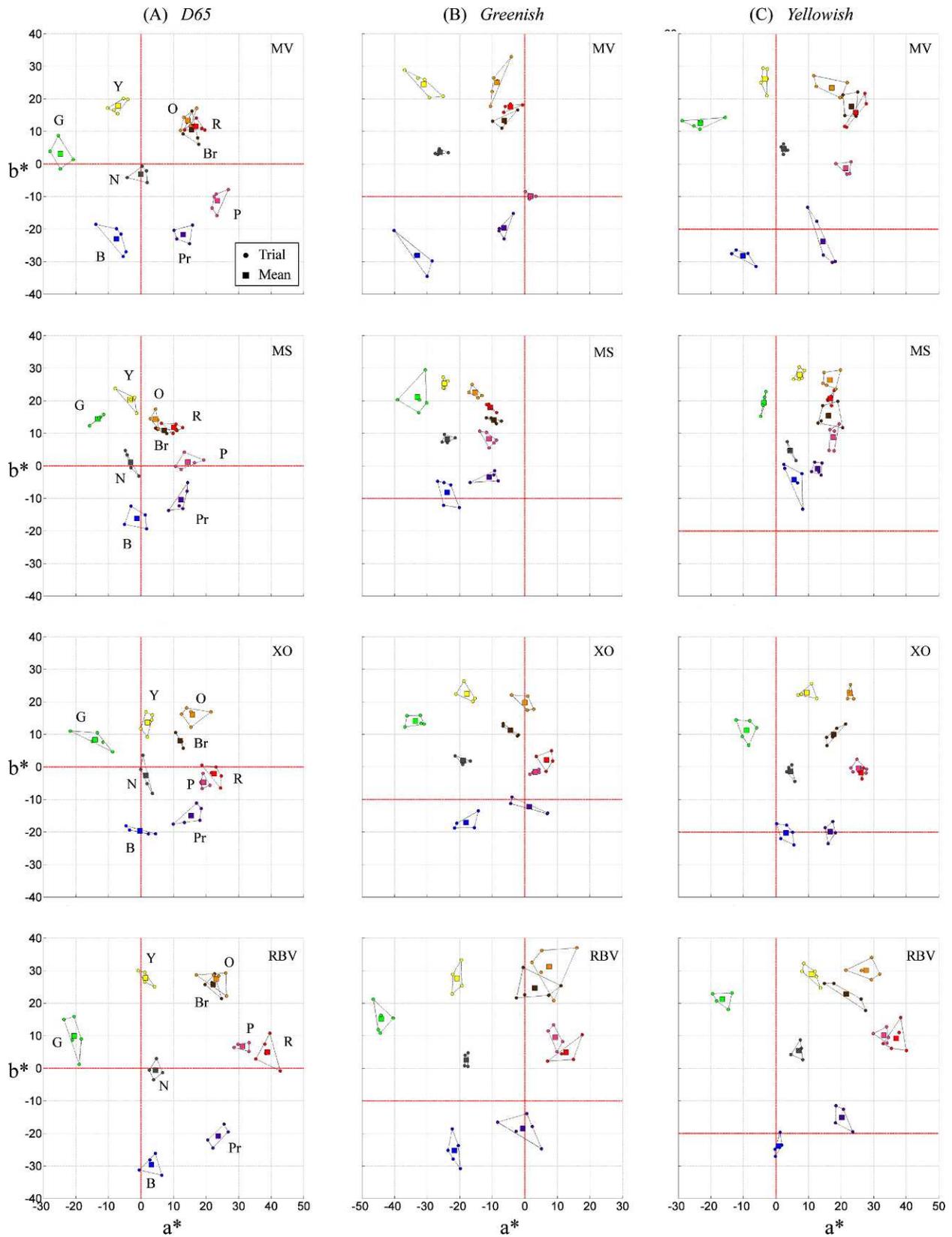
Figure 7. Individual chromatic settings results from four different subjects (regular sessions). Column A: under D65 illuminant; column B: under greenish illuminant; and column C: under yellowish illuminant. The background was Type II in all cases. Individual trials are represented by small dots joined by lines and their average is represented by a color-coded square.

	Red	Green	Blue	Yellow	Neutral	Purple	Pink	Orange	Brown	Mean
JRV	2.14	3.82	3.44	3.42	2.35	3.87	3.26	3.42	2.60	3.18
CAP	4.40	3.56	3.53	3.45	3.81	5.96	4.66	5.75	4.96	4.45
MV	3.51	4.31	5.60	5.10	2.91	6.62	4.66	5.25	4.34	4.70
MS	2.44	4.05	5.82	4.65	3.39	3.78	5.43	3.37	2.78	3.97
XO	2.98	3.87	3.33	3.28	3.00	4.38	3.27	4.02	3.22	3.48
RB	3.65	3.51	3.61	3.42	7.39	4.44	3.62	4.71	6.70	4.56
LC	5.94	6.55	5.06	7.61	4.71	5.76	5.87	4.85	7.11	5.94
AB	4.29	5.66	5.23	4.34	5.09	6.55	5.55	5.21	4.47	5.15
RBV	5.17	5.20	4.93	5.06	3.08	4.87	4.94	6.71	5.14	5.01
JC	5.54	4.45	5.67	4.94	4.81	6.63	6.19	6.92	4.79	5.55
Mean	4.03	4.50	4.62	4.53	4.05	5.29	4.74	5.02	4.61	**4.60**

Table 3. Variability (δ) of mean chromatic settings in ΔE^* units, averaged over illuminants and backgrounds. The columns show values according to color category and the rows according to subject. The last column/row shows the means of the rows/columns. The value in bold corresponds to the overall mean.

operate in LMS cone excitation coordinates (Brainard et al., 1997; Burnham, Evans, & Newhall, 1957; Jameson & Hurvich, 1964), calculated from the Smith and Pokorny cone sensitivity functions (Smith & Pokorny, 1975). Equation 2 formalizes the previous approach where **x** and **y** are the LMS cone excitations produced by the light reaching the observer from the CRT monitor: **x** corresponds to the reference illuminant (D65) and **y** corresponds to the test illuminant (greenish or yellowish).

$$\mathbf{y} = \mathbf{M}[\mathbf{x}\,1]^{\mathrm{T}} \quad \text{where}$$

$$\mathbf{M} = \begin{pmatrix} m_{1,1} & m_{1,2} & m_{1,3} & m_{1,4} \\ m_{2,1} & m_{2,2} & m_{2,3} & m_{2,4} \\ m_{3,1} & m_{3,2} & m_{3,3} & m_{3,4} \end{pmatrix} \in \mathbb{R}^{3 \times 4} \quad (2)$$

The model is represented by the matrix **M,** which can take one of several possible forms according to its nonzero coefficients. These can also be understood in terms of models of visual mechanisms:

Diagonal (D)

The diagonal model ($m_{i,j} = 0$ if $i \neq j$) has only three free parameters. This model only allows for multiplicative gain changes that are specific to each one of the three cone classes. It is often referred as Von Kries adaptation (Brainard & Wandell, 1992; Von Kries, 1905/1970).

Linear (L)

The linear model ($m_{i,j} = 0$ if $j = 4$) has nine free parameters. This model allows signals from each cone type to be modulated independently and can describe multiplicative gain changes both at the receptor level

Category/Index	CI		$\overline{BR}$		CCI		
Category/Illuminant	Greenish	Yellowish	Greenish	Yellowish	Greenish	Yellowish	Mean
Red	**0.37**	0.63	0.69	0.65	0.82	0.76	0.65
Green	**0.73**	0.68	0.61	0.58	0.89	**0.88**	0.73
Blue	0.53	**0.55**	0.64	0.65	0.68	0.68	0.62
Yellow	0.71	**0.76**	**0.51**	**0.49**	0.72	0.75	0.66
Gray	0.55	0.56	0.62	0.63	**0.61**	**0.62**	0.60
Purple	0.49	0.58	0.72	**0.78**	0.77	0.79	0.69
Pink	0.55	0.64	0.54	0.58	0.64	0.68	0.60
Orange	0.62	0.75	0.53	0.51	0.73	0.76	0.65
Brown	0.50	0.70	**0.75**	0.57	**0.96**	0.82	0.72
Mean	0.56	0.65	0.62	0.60	0.76	0.75	0.66

Table 4. Three color constancy indices applied to our measures and split by color categories and illuminant type. *Notes*: All indices were computed in the CIE1976 UCS uv uniform color space and averaged for all subjects and backgrounds. We highlighted in bold the maximum and minimum values in each column, which reveal considerable differences within color categories.

and after an opponent transformation (Brainard & Wandell, 1992).

Affine (A)

The affine model does not set any initial coefficient to zero, and it has 12 free parameters. It contains nested versions of the previous two models. The first three columns of $\mathbf{M}$ include the linear model and the fourth column represents an additive process. This model can be thought as an instance of the two-process model proposed by Jameson and Hurvich (1964) (see also Brainard & Wandell, 1992).

Diagonal plus Translation (DT)

The diagonal plus translation model ($m_{i,j} = 0$ if $i \neq j$ and $j < 4$) has six free parameters and can be seen as a simplification of the affine model. The first three columns allow only for multiplicative gains for each cone class and the last column allows a further additive process.

We studied the predictive power of each model when multiple chromatic settings were used as data points. Equation 3 generalizes Equation 2 into a single system of linear equations when using more than one data point. This formulation allows using standard multiple linear regression methods to fit the model parameters, i.e., to minimize the mean-square difference between the measured and the predicted points. In Equation 3, the matrix $\mathbf{X}$ contains the LMS coordinates of n colors $\mathbf{x}_i$ under reference illuminant and matrix $\mathbf{Y}$ contains the settings of those same colors, $\mathbf{y}_i$, under test illuminant.

$$\mathbf{Y} = \mathbf{MX} \quad \text{where}$$
$$\mathbf{Y} = (\mathbf{y}_1 | \cdots | \mathbf{y}_n) \in \mathbb{R}^{3 \times n} \quad \text{and}$$
$$\mathbf{X} = ([\mathbf{x}_1\, 1]' | \ldots | [\mathbf{x}_n\, 1]') \in \mathbb{R}^{4 \times n} \quad (3)$$

Equation 4 describes $\mathcal{H}$, which contains all possible subsets of nine colors and their combinations according to their indices (one for green, two for blue, three for yellow, etc.). Once a particular element of $\mathcal{H}$ was selected we could fit the model parameters to this element as described in Equation 5, substitute their LMS coordinates and solve the linear system using least squares. However, since LMS is not perceptually uniform, we chose to follow the approach described by Brainard and colleagues (Brainard et al., 1997; Brainard & Wandell, 1992). They solved the linear system through a minimization process which determined the model parameters according to the mean CIELab ΔE^* color difference between the N predictions and the data points. The function to minimize is described by Equation 6, where φ is an operator that translates from LMS to CIELab coordinates. F_N was minimized using the Matlab Optimization Toolbox. Model precision was

evaluated by computing the average ΔE^* difference between the whole set of nine chromatic settings and their predictions computed from the matrix $\mathbf{M}$.

$$\mathcal{H} = \{(k_1, \ldots, k_N); \, k_i \in \{1, \ldots, 9\}, \, k_i \neq k_j,$$
$$N = 1, \ldots, 9 \, \text{ and } \, i, j = 1, \ldots, N\} \quad (4)$$

$$(\mathbf{y}_{k_1} | \cdots | \mathbf{y}_{k_N}) = \mathbf{M}([\mathbf{x}_{k_1}\, 1]' | \cdots | [\mathbf{x}_{k_N}\, 1]')$$
$$\text{where} \quad (k_1, \ldots, k_N) \in \mathcal{H} \quad (5)$$

$$F_N(\mathbf{M}, (k_1, \ldots, k_N)) = \frac{1}{N} \sum_{i=1}^{N} \left(\varphi(\mathbf{M}[\mathbf{x}_{k_i}\, 1]') - \varphi(y_{k_i}) \right)_2$$
$$\text{where} \quad \varphi(LMS) \in CIELab \quad (6)$$

We considered all possible combinations of SRs, within the limits imposed by each model. For example, when fitting the linear system in Equation 5, the minimum number of points that the model can fit is determined by the number of free parameters contained in the model. This terminology is equivalent to a system of linear equations where there are larger, fewer, or equal number of equations than unknowns. The underdetermined case occurs when the number of unknowns is larger than the number of the equations (the system is underconstrained). From this follows that the diagonal model admits any number of data points $N \geq 1$, diagonal plus translation admits $N \geq 2$ data points, linear admits $N \geq 3$ data points, and affine $N \geq 4$ data points. This is also valid for Equation 6.

Figure 8 summarizes our modeling results as described above. Panel A corresponds to greenish illuminant and panel B to yellowish. The y-axis shows the prediction error (in ΔE^* units) associated with each model as a function of the number of chromatic settings used to fit it. Following the approach of Brainard et al. (1997), we used the chromaticity coordinates of the corresponding illuminant as a reference white point in each case. The function specified in Equation 6 was minimized to fit chromatic settings $\mathbf{x}$ (corresponding to D65) and $\mathbf{y}$ (corresponding to greenish or yellowish illuminants) keeping the same background type. Take for instance panel A in Figure 8, where each point is the average model prediction error from all possible combinations of elements of $\mathcal{H}$ that contain the number of colors specified in the x-axis, across backgrounds and subjects. Consider the case when the nine SRs were measured both under D65 and greenish illumination using the same background type. We fitted the diagonal model to only one correspondence pair from the nine chromatic settings available and used the same parameters to predict the positions of all nine corresponding pairs. We repeated this for all the other pairs and calculated the average
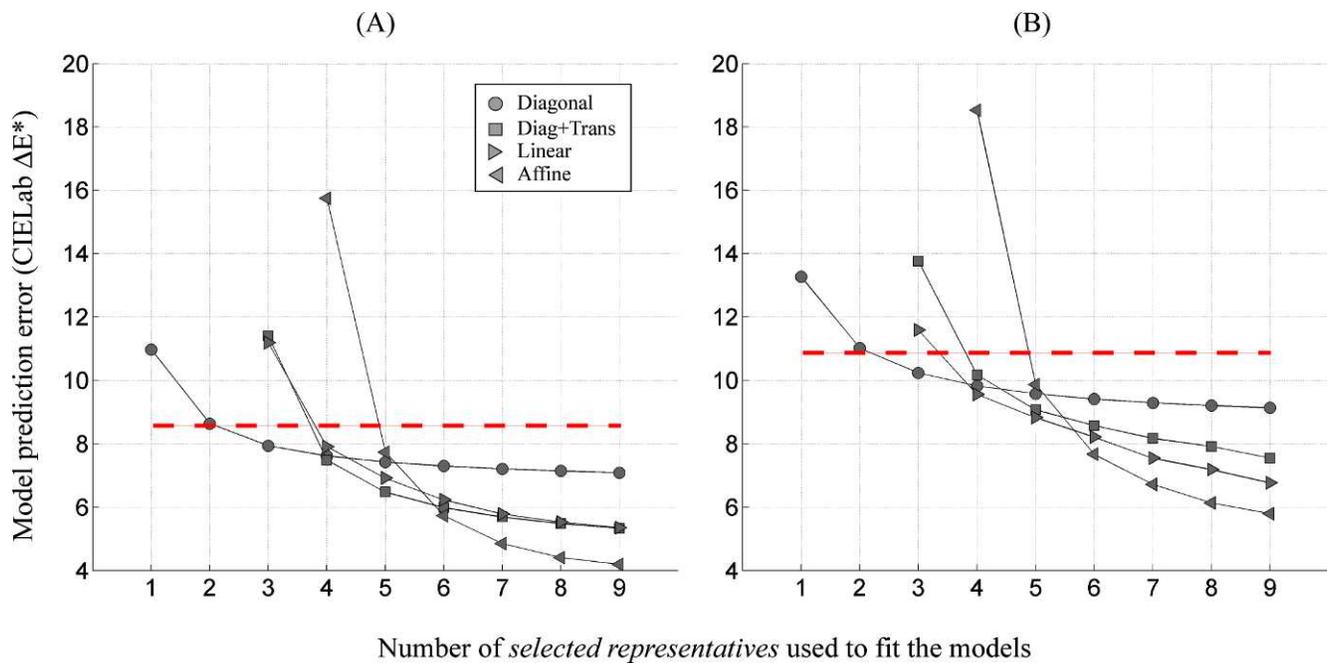
Figure 8. Model prediction error according to the number of colors used to estimate their parameters. Panel A corresponds to greenish test illuminant and panel B to yellowish. Each point corresponds to a particular model (circles for the Diagonal, squares for the Diagonal plus Translation, right-pointing triangles for the Linear, and left-pointing triangles for the Affine), computed from all background types and subjects. For comparison we show the prediction error of Von Kries transformation applied to the achromatic locus as a horizontal red broken line. The values were calculated using the corresponding reference white point for each illuminant (greenish and yellowish—see Table 1).

CIELab ΔE^* distance between predicted and measured points for the nine chromatic settings pairs. We extended this to all subjects and backgrounds. The result of these calculations (average from 270 model predictions) is shown in panel A as the leftmost filled circle in the plot. To calculate the second leftmost circle in the plot, we fitted the diagonal model to two correspondence pairs from the nine chromatic settings available and predicted the positions of all nine pairs (36 possible combinations). This point represents the average across subjects and backgrounds (1,080 model predictions). The other circles were calculated similarly by fitting the diagonal model to increasingly more data points. The same reasoning was applied to the other models, shown as triangles and squares in Figure 8. Since the results of the minimization process in Equation 6 depend on the initial seed, we used 100 random seeds (for larger values results tend to stabilize) and the solution to the linear system specified by Equation 5 (Brainard & Wandell, 1992) as a complementary seed. We selected the minimum optimization value of all seeds.

Predictably, Figure 8 shows that adding more data points and increasing the number of free parameters lowers the model prediction error exponentially: the more free parameters a model has, the more accentuated the decay is. For instance, the Diagonal model (circle symbols) improves less, from 10.9 to 7.1 ΔE^* for

the greenish and from 13.3 to 9.1 ΔE^* for the yellowish as we add more fitting points. When the maximum number of fitting points (nine) are used, the errors in ΔE^* are: 7.09 (D), 5.33 (DT), 5.35 (L), and 4.19 (A) for the greenish illuminant and 9.13 (D), 7.55 (DT), 6.76 (L), and 5.79 (A) for the yellowish illuminant (see Figure 9). In general, model errors under greenish illuminant are lower than model errors under yellowish illuminant. Simpler models tend to perform better with a small number of fitting points whereas more complex models tend to perform better with larger numbers of fitting points. For instance the Linear and Affine models start to perform better than the simpler Diagonal when more than five points are considered. There are also quantitative differences regarding the illuminant: For up to five fitting points, error values are between 4 and 7.5 ΔE^* for greenish and between 6 and 9.6 ΔE^* for the yellowish.

We tested the parsimony of the models to see whether they include more parameters than it is necessary by applying the *Akaike Information Criterion* (Burnham & Anderson, 2002). This criterion measures the relative goodness of fit of a model in terms of the information lost when it is used to describe data (see Appendix B). The results show that the best models in Figure 8 are the simplest: Diagonal and Diagonal plus Translation, implying that the Linear and the Affine models are possibly over-fitting the data. In particular,
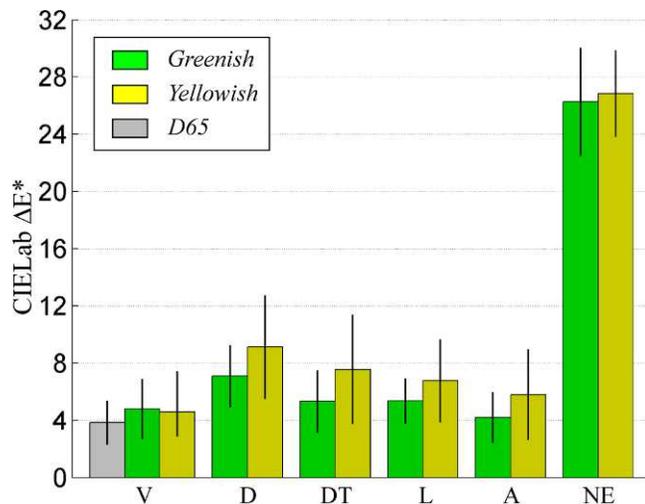
Figure 9. Models' prediction error when all nine SRs points were included. The first column (V) contains the subject average variability in the trials (see Table 3). The last column (No Effect − NE) is a quantitative measure of the illuminant shift computed using the Identity matrix as a model. The groups of bars labeled as D, DT, L, and A correspond each to the Diagonal, Diagonal plus Translation, Linear, and Affine models, respectively.

Akaike weights, which indicate the plausibility for each model being the best are equal to zero for the Linear and Affine. The results also show a clear tendency for the Diagonal plus Translation to become the best in terms of number of free parameters and prediction error as we add more data points (*Akaike weights* increase with increasing number of data points for the Diagonal plus Translation model).

Discussion

Our previous results show the feasibility of using several colors rather than a single color as a metric for assessing the stability of color appearance under a change of illumination. In the following section we discuss the usefulness of this new metric, showing that linear color constancy models satisfactorily explain the transformations with a larger number of colors. At the end of the section we introduce a new color constancy index that takes into account several aspects of color constancy not considered before.

Does including more colors increase the precision of linear color constancy models?

Both graphs in Figure 8 illustrate clearly how the predictive power of all models is increased by adding more fitting points, something that is in agreement with previous studies (Hansen et al., 2007; Olkkonen et al., 2009; Olkkonen et al., 2010). However, the error curves tend to a constant value after eight SRs, and this suggests that measuring more points would lead to minimal improvements. In this context, it is worth noticing that our current fitting points were not determined randomly but had a particularly even distribution over the color space; thus our conclusions become more relevant when all nine fitting points are used. This highlights the advantage of measuring several colors instead of just gray, and although it disagrees with previous results (Speigle & Brainard, 1999), we believe it is unlikely to be the product of experimental artifacts. Figure 9 shows the portion of the phenomenon that is captured by the models. The large differences in height between the bar labeled as "No-Effect" (which summarizes the effects of the illumination), and the other bars suggest that all linear models succeed in modeling the phenomenon (Brainard, 1998; Brainard et al., 1997). However there is still a small part that is not captured by the models.

If we ignore the Linear and Affine models, in Figure 8 there are some common qualitative features for both illuminants that are worth mentioning:

Stability point at five SRs

All models approximately have the same precision when five SRs are used for the fit, i.e., three free parameters achieve similar results as twelve. This might reflect the fact that considering less than five points in our calculations allows for distributions of colors that are not symmetric with respect to the center, something that is less likely when more colors are considered. Furthermore, models with more free parameters are more sensitive to these asymmetries.

Diagonal outperforms the Diagonal plus Translation before the stability point

This suggests a link between the number of colors available and the complexity of the color constancy mechanism needed: In a simpler environment, a cone gain-based transformation outperforms the others.

Diagonal plus Translation outperforms the Diagonal after the stability point

This represents an improvement from the Diagonal model and suggests the involvement of the additive process in a two-stage mechanism as proposed by Jameson and Hurvich (1955).

Interestingly, the modeling of the chromatic settings performed under the greenish illuminant is better than under the yellowish one, and this effect is general to all

models and fitting point numbers. This fact suggests a higher degree of dispersion in the chromatic settings, which may result from the split of the resulting colors into several categories when illuminated by the yellowish illuminant, something that did not occur under the greenish illuminant (see further discussion below).

Further insights into the role of color categories

The overall pattern of results shown in the previous sections is broadly uniform across color categories, but some particularities exist. For example, we expected the behavior of gray (the color measured in achromatic settings) to be outstanding in terms of variability (δ), adjustment time, and constancy index and to summarize the behavior of the whole set of chromatic settings. Interestingly, we have found that subject's ability to adjust gray and red are similar, closely followed by many other categories. Also, gray is the color that takes longer time to adjust, maybe because subjects can discriminate more finely near the achromatic locus (Boynton & Olson, 1987). Furthermore, we expected color constancy indices for gray to be near the average, and Table 4 shows that they are generally low, and in the case of the CCI index, the lowest. Previous work found higher color constancy for gray than for chromatic stimuli (Olkkonen et al., 2010; Speigle & Brainard, 1999), which is perhaps due to the fact that we used simulated surfaces and illuminants instead of real surfaces. We also found high color constancy values (0.66 in average), which is in accordance to similar studies (Foster, 2011; Hansen et al., 2007; Ling & Hurlbert, 2008; Murray, Daugirdiene, Vaitkevicius, Kulikowski, & Stanikunas, 2006; Olkkonen et al., 2009; Olkkonen et al., 2010), a fact that is supported by visual inspection of the plots in Figure 6, where interdistances among measured colors are largely preserved. This supports the finding that the categorical structure of color space is largely preserved under illuminant changes (Hansen et al., 2007; Olkkonen et al., 2009; Olkkonen et al., 2010).

The differences in color constancy values found for different categories suggest different properties for different categorical colors. This implies that we should always refer to the same color category when comparing across color constancy measures. In the case where we pool measures across several categories to produce a single color constancy index (see below), these color categories should be maintained for the index to be consistent.

The particular properties of each categorical color in terms of color constancy indices are likely determined by stimulus configurations and subjects' tasks. However, here we found no differences in terms of

Property/Index	CI, BR, $\overline{BR}$	EI	CCI	SCI
Magnitude	Yes	Yes	Yes	Yes
Orientation	No	No	Yes	Yes
Memory	No	Yes	Yes	Yes
Structure	No	Yes	No	Yes

Table 5. Summary of some properties of color constancy incorporated into each index.

background types, a result which is similar to others (Brainard, 1998).

SCI: A new structural color constancy index

Color constancy indices attempt to capture the extent of the phenomenon's effect in a single number. They relate perceptual data measured under a state of adaptation to the corresponding data predicted for "perfect" adaptation (i.e., physical color shift). The simplest indices quantify Euclidean distances (*magnitude*) among the colors of the test surface, the ideal match and the observer match. Examples of these are the CI (Arend et al., 1991), the BR (Troost & de Weert, 1991), and the BRϕ, which incorporates the direction (*orientation*) between the perceptual and physical color shifts (Foster, 2011). Several improvements have been suggested. For instance, Ling and Hurlbert (2008) proposed a new index CCI that incorporates the matching error in the absence of illumination change (*memory* shift) and Brainard (1998) proposed to use the Equivalent Illuminant (EI) instead of the measured adaptation point, which is calculated from different measured points and thus captures the inter-distances among the colors considered under a given adaptation state (*structural*).

Following the previous discussion, we introduced a new color constancy index, termed S*tructural Constancy Index* (SCI), which captures all the features stated in Table 5. The new index is defined in terms of matrix norms, which are extensions of the notion of vector norms applied to matrices. As Equation 7 shows, the norm of a matrix **A** is obtained from the norm of vectors **x** and **Ax** and describes the maximum relative vector magnitude change under the linear transformation **A**.

$$||\mathbf{A}||_2 = \sup_{\mathbf{x} \neq 0} \frac{||\mathbf{Ax}||_2}{||\mathbf{x}||_2} = \max_{(\mathbf{x})_2 = 1} ||\mathbf{Ax}||_2 \qquad (7)$$

In our context, the matrix **A** models the effects of the illuminant change, i.e., given the coordinates **x** of a color sample under the reference illuminant, it returns the coordinates **Ax** of the same sample under the test illuminant in a given color space. We define SCI as:

$$SCI(\mathbf{A_{percep}}, \mathbf{A_{phys}}) = \frac{||\mathbf{A_{percep}}||_2}{||\mathbf{A_{phys}}||_2} \cdot \cos(angle(\mathbf{r}, \mathbf{s}))$$

$$= \frac{||\mathbf{A_{percep}}||_2}{||\mathbf{A_{phys}}||_2} \cdot \frac{\mathbf{rs}}{||\mathbf{r}||_2 ||\mathbf{s}||_2}$$

$$(8)$$

In Equation 8 SCI is defined as the product of two factors. The first factor is the quotient of two matrix norms, and computes the relative magnitudes of the perceptual and physical effects of the illuminant, as is commonly the case with constancy indices (Arend et al., 1991; Foster, 2011; Hansen et al., 2007; Ling & Hurlbert, 2008; Smithson & Zaidi, 2004; Yang & Shevell, 2002). The second factor estimates how much the direction of the adaptation coincides with the direction of the actual illuminant change in the color space considered. To compute this we need both $\mathbf{A_{percep}}$ and $\mathbf{A_{phys}}$ to be affine matrices, with the last column of $\mathbf{A_{percep}}$ specifying the translation vector $\mathbf{r}$ and the last column of $\mathbf{A_{phys}}$ specifying the translation vector $\mathbf{s}$. Notice that here we are not using an affine matrix to model the data as in previous sections, but to quantify two different aspects of color constancy: magnitude and orientation.

The coefficients of $\mathbf{A_{percep}}$ are determined from pairs of corresponding chromatic settings under reference and test illuminants and can be obtained following the approach described in the modeling subsection above (Equation 3). Likewise, the coefficients of $\mathbf{A_{phys}}$ are determined from correspondences between the chromatic settings made under the reference illuminant and physical simulations of the same colors under a test illuminant. In this formulation, if matrices $\mathbf{A_{percep}}$ and $\mathbf{A_{phys}}$ are equal, then color constancy is perfect. Finally, memory effects like those discussed by Ling and Hurlbert (2008) are neutralized since our measurements were obtained from direct comparisons under reference and test illuminants.

Figure 10 illustrates the behavior of Equation 8 for several hypothetical cases. Panel A describes how the magnitude size of each transformation contributes to the value of the SCI. This contribution is always positive and can be smaller or larger than one according to the ratio between the norms of the $\mathbf{A_{percep}}$ and $\mathbf{A_{phys}}$. The latter case happens when observers correct for the illuminant more than they should. Panel B describes the contribution of the second term of Equation 8, i.e., a weighting factor to penalize for angular deviations from the direction of the simulated illuminant shift. As $\mathbf{r}$ and $\mathbf{s}$ become more perpendicular, their product $\mathbf{rs}$ becomes closer to zero. Although negative values are possible in theory, in practice this weighting factor should be positive assuming that $\mathbf{r}$ and $\mathbf{s}$ are far from perpendicular. Structural information of the color constancy phenomenon is implicitly embedded in the affine matrix.

Other indices such as BR would produce the same value for all hypothetical settings located around the half circumference defined by the broken line, since it only compares the magnitudes of both shifts. For the hypothetical cases described by $\mathbf{s}_1$ and $\mathbf{s}_3$, CI would have a value of one, since it compares the magnitude of the vector defined by the settings and their expected location to the magnitude of the illuminant shift. Panels C and D illustrate how structural information is summarized into a single positive number. Popular indices such as CI, BR, and CCI do not convey this information since they are usually computed over the achromatic setting. Panel C illustrates the case when there is no translation (i.e., the last column of the affine matrix is null) and the matrix can be interpreted in terms of expansion ($||A||_2 > 1$), retraction ($||A||_2 < 1$), or rotation ($||A||_2 = 1$). Panel D illustrates the case when only the translation part is operative and the value of the norm reflects this translation. Panel E shows an exemplary case when the spatial relationships among measurements are disrupted by just one chromatic setting *outlier*. This structural disruption would be embedded in the affine matrix, which represents a compromise solution in between the prediction error of the outlier and the rest of measurements. Notice, that in this particular example there is only one outlier but there could be more, with effects such as those described in Panel C (contraction, expansion, and rotation), leading to more complex outcomes. As Table 4 shows, each index produces a different value for a different color category; SCI deals with this variability by summarizing the measurements for all categories into an affine matrix. This casuistic is illustrated by panel E whose chromatic setting outlier would have produced values of CI, BR, or SCI notably different from the rest, thus making the quantification of color constancy dependent on the selected color category.

In common with other indices, the SCI can, in theory, assume values that are larger than one or negative, representing overcompensation or failures of color constancy that may happen under certain illumination conditions such as multiple illuminants, non-Lambertian surfaces, self-luminous or fluorescent materials, etc., that imply a violation of the initial conditions of this analysis.

Table 6 shows the average values obtained from applying four color constancy indices ($\overline{BR}$, EI, CCI, and SCI) to all subjects and background types, discriminated according to illumination. All indices were computed in the CIE1976 uv color space (the SCI index was computed using the affine matrix described above). There was no effect of background types in the calculations. Interestingly not all indices gave the same values; EI and $\overline{BR}$ were generally lower and SCI was the highest. The differences between popular indices such as $\overline{BR}$ and CCI were reported by Ling and

(A) Magnitude ($o = 1$)

$$\left\|\mathbf{A}_{\mathbf{percep}}\right\|_2 > \left\|\mathbf{A}_{\mathbf{phys}}\right\|_2 \Rightarrow \mathrm{SCI} > 1$$

$\left\|\mathbf{A}_{\mathbf{phys}}\right\|_2$

$$\left\|\mathbf{A}_{\mathbf{percep}}\right\|_2 = \left\|\mathbf{A}_{\mathbf{phys}}\right\|_2 \Rightarrow \mathrm{SCI} = 1$$

$$\left\|\mathbf{A}_{\mathbf{percep}}\right\|_2 < \left\|\mathbf{A}_{\mathbf{phys}}\right\|_2 \Rightarrow 0 < \mathrm{SCI} < 1$$

(B) Orientation ($m = 1$)

$$\mathrm{SCI} = \cos\left(angle\left(\mathbf{r}, \mathbf{s}_3\right)\right)$$

$$\mathbf{r} \parallel \mathbf{s}_2 \Rightarrow \mathrm{SCI} = 1$$

$\mathbf{s}_3$ $\mathbf{r}$ $\mathbf{s}_2$

$\mathbf{s}_4$

$\mathbf{s}_1$

$$\mathbf{r} \perp \mathbf{s}_1 \Rightarrow \mathrm{SCI} = 0$$
$$\mathbf{r} \perp \mathbf{s}_4 \Rightarrow \mathrm{SCI} = 0$$

(C) Structure (no translation; $\left\|\mathbf{A}_{\mathbf{phys}}\right\|_2 = 1$; and $o = 1$)

$$\mathrm{SCI} = \left\|\mathbf{A}_{\mathbf{percep}}\right\|_2 < 1$$

$$\mathrm{SCI} = \left\|\mathbf{A}_{\mathbf{percep}}\right\|_2 = 1$$

$$\mathrm{SCI} = \left\|\mathbf{A}_{\mathbf{percep}}\right\|_2 > 1$$

(D) Structure (only translation (t_1, t_2); $\left\|\mathbf{A}_{\mathbf{phys}}\right\|_2 = 1$; and $o = 1$)

$(t_1 \neq 0, t_2 = 0)$ $(t_1 > 0, t_2 > 0)$

$$\mathrm{SCI} = \left\|\mathbf{A}_{\mathbf{percep}}\right\|_2 = 1 + |t_1| \quad\quad \mathrm{SCI} = \left\|\mathbf{A}_{\mathbf{percep}}\right\|_2 > 1$$

$(t_1 = 0, t_2 = 0)$ $(t_1 = 0, t_2 \neq 0)$

$$\mathrm{SCI} = \left\|\mathbf{A}_{\mathbf{percep}}\right\|_2 = 0 \quad\quad \mathrm{SCI} = \left\|\mathbf{A}_{\mathbf{percep}}\right\|_2 = 1 + |t_2|$$

u'

(E) Structure disruption (no translation)

$$\mathrm{SCI} = \left\|\mathbf{A}_{\mathbf{percep}}\right\|_2 > 1$$

$$\mathrm{SCI} = \left\|\mathbf{A}_{\mathbf{percep}}\right\|_2 < 1$$

u'

$$\mathrm{SCI} = magnitude \cdot orientation = m \cdot o$$

$$m = \left\|\mathbf{A}_{\mathbf{percep}}\right\|_2 \Big/ \left\|\mathbf{A}_{\mathbf{phys}}\right\|_2$$

$$o = \cos\left(angle\left(\mathbf{r}, \mathbf{s}\right)\right)$$

○ Physical colors
□ Hypothetical settings under test illum
■ Settings under reference illum

Figure 10. Hypothetical cases of chromatic settings and their contribution to SCI values in the CIE1976 uv color space. Each panel illustrates the contribution of a particular feature of our index. Dark squares correspond to chromatic settings made under the reference illuminant, light squares correspond to hypothetical chromatic settings made under test illuminant, and circles correspond to a simulated illumination of the chromatic settings made under the reference illuminant (dark squares - physical colors). The broken lines group hypothetical chromatic settings done under the same illuminant condition. Panel A: effects of a shift in magnitude only with respect of a simulated illumination. Panel B: effects of a change in the orientation from the simulated illuminant shift. Panels C and D: effects of an expansion/contraction and a translation are captured and converted into a single number by the affine matrix norm. Panel E: effects of the different spatial relationships on the matrix norm. Notice the different constraints detailed in the tiles of Panels C, D, and E for each SCI example.

Hurlbert (2008) and attributed to the incorporation of memory shift into the index formula. SCI values are slightly higher than CCI values, and in the case of the greenish illuminant, larger than one. This fact is due to the incorporation of "structural" components, i.e., measures of the interdistances among data points into the index calculation (see panel C in Figure 10), which can increase the total index value in some cases. We calculated the contribution of the different components in Figure 10 to the SCI values in Table 6 and found that, for the greenish illuminant, the norm of $\mathbf{A_{percep}}$ is slightly larger than the norm of $\mathbf{A_{phys}}$, making the first term of Equation 8 slightly larger than one. The previous analysis implies that perfect color constancy is achieved when SCI is equal to one and different values indicate either lack of constancy (SCI <1) or overcompensation (SCI >1). In our case, we expected values close to one due to the long adaptation period of immersive illumination.

We tested whether the high values we found in Table 6 were due to the fact that observers had the chance to see the Type I background colors (i.e., the colors to be adjusted) often, and hence subjects performed matches to the displayed colors instead of reproducing them from their memory. This was done by repeating the experiment with two new subjects using only Type II background, i.e., they had not seen the Type I backgrounds colors before (see Appendix C). Their results were in agreement with those of the rest of the subjects, and indeed their color constancy indices were not lower than those of Table 6.

Table 6 reveals that only SCI differentiates between the greenish and the yellowish illuminants. Further inspection of the *magnitude* and *orientation* contributions revealed that these differences originated in the norm of the perceptual matrix as explained in Panels C, D, and E of Figure 10. In the previous modeling subsection, we found lower prediction errors for the greenish illuminant (see Figure 8), indicating that such data is better captured by the fitting of linear models, a process similar to the computation of SCI values. This explains why chromatic settings under yellowish illuminant have a higher degree of dispersion when compared to chromatic settings under D65 than in the greenish case. These differences manifest in Figure 6 as subtle variations in the location of the yellow, orange, brown, red, and pink data points, which may account for the 18% difference between both illuminants in Table 6. We could hypothesize about the origin of this dispersion and say that greenish-illuminated colors fall inside the broad green category, whereas yellowish-illuminated colors fall into several categories and this initial (first milliseconds) categorical perception may influence the subject's adaptation and subsequent chromatic settings. However, this needs to be settled by doing more experiments in the future.

Index/Illuminant	Greenish	Yellowish
BR	0.62	0.61
EI	0.58	0.59
CCI	0.76	0.75
SCI	1.03	0.85

Table 6. The *Structural Constancy Index* (SCI) and other typical color constancy indices computed in the CIE1976 uv. Each value corresponds to the average over subjects and background types, also for the CCI and EI averaged over color categories.

Comparison to previous paradigms

Our contribution is complementary to the work of others who have also studied successive color constancy (Foster, 2011) under large periods of immersive illumination and have used simulated (Hansen et al., 2007; Olkkonen et al., 2009) or real (Olkkonen et al., 2010) surfaces. These studies categorize a large number of colored samples with higher results variance, while we measured only nine relevant points with relatively higher precision. Hansen et al. (2007; Olkkonen et al., 2009) measured changes in the categorical boundaries of the color space while Olkkonen et al. (2010) used a conventional constancy index (including shift magnitude and orientation) applied to the categorical prototypes. Our results qualitatively agree with their findings regarding the stability of the categorical structure of color space under illuminant changes (Hansen et al., 2007; Olkkonen et al., 2009; Olkkonen et al., 2010).

The chromatic setting paradigm was primarily designed to deal with two main issues: (a) the state of adaptation closely following the change of illuminant (Foster, 2011) and b) the effects of instructions regarding the nature of the stimuli (surface-match or color-match criteria) (Troost & de Weert, 1991). For this reason, it makes use of subject's color naming abilities, asking them to select their own colors instead of reproducing arbitrary ones, thus improving on the chromatic resolution limits of standard color-naming techniques (Foster, 2011). The main disadvantage of the method is arguably the saturation restriction to the colors that subjects can initially select imposed by the CRT monitor gamut limitations. However low-saturation SRs were not particularly difficult to reproduce in regular sessions.

Possible chromatic induction (Shevell & Wei, 2000) effects resulting from the local influence of neighboring patches were avoided by embedding the multiple test patch within the Mondrian, randomizing its spatial and chromatic structure from trial to trial (while keeping its global statistics constant prior to illumination). In this manner, subjects have to look at several places and average the test patch color before making a decision.

Also, general memory effects (Ling & Hurlbert, 2008) were isolated from constancy effects by analyzing memory matches with and without the illuminant change.

Conclusions

We developed a new paradigm (the *Chromatic Setting*) to study color constancy, which measures several points in color space under extended periods of adaptation to the illumination. We have shown the paradigm to be feasible in terms of memory and consistency of subject's responses over time. No remarkable differences were found between the role played by gray and the rest of the chromatic categories tested for this task. Our results show that linear models, in particular the Diagonal plus Translation, succeed in capturing the color constancy phenomenon. They also show that including more colors does improve model precision. A quantification of the phenomenon in terms of commonly used color constancy indices reveals substantial differences when applied to individual colors. In addition to our paradigm, we developed a more comprehensive color constancy index (the *Structural Constancy Index*), which accounts for changes in magnitude, orientation, and structure, as well as memory effects. When applied to our measures, our index indicates nearly full constancy for the greenish illuminant and slightly less constancy for the yellowish illuminants tested. Our results do not show any quantitative difference regarding the types of colored background tested.

Keywords: color vision, color appearance/constancy, categorization, computational modeling

Acknowledgments

Commercial relationships: none.
Corresponding author: Jordi Roca-Vila.
Email: jroca@cvc.uab.cat.
Address: Computer Vision Center, Bellaterra, Barcelona, Spain.

References

Arend, L., & Reeves, A. (1986). Simultaneous color constancy. *Journal of the Optical Society of America A: Optics, Image Science, & Vision, 3*(10), 1743–1751.

Arend, L., Reeves, A., Schirillo, J., & Goldstein, R. (1991). Simultaneous color constancy—Papers with diverse Munsell values. *Journal of the Optical Society of America A: Optics, Image Science, & Vision, 8*(4), 661–672.

Benavente, R., Parraga, C. A., & Vanrell, M. (2009). Colour categories boundaries are better defined in contextual conditions. *Perception, 38*(Suppl.), 36.

Berlin, B., & Kay, P. (1991). *Basic color terms: Their universality and evolution.* Berkeley, CA: University of California Press. (Original work published 1969.)

Boynton, R. M., & Olson, C. X. (1987). Locating basic colors in the OSA space. *Color Research & Application, 12*(2), 94–105.

Brainard, D. H. (1998). Color constancy in the nearly natural image. II. Achromatic loci. *Journal of the Optical Society of America A: Optics, Image Science, & Vision, 15*(2), 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: Optics, Image Science, & Vision, 14*(9), 2091–2110.

Brainard, D. H., & Wandell, B. A. (1986). Analysis of the retinex theory of color vision. *Journal of the Optical Society of America A: Optics, Image Science, & Vision, 3*(10), 1651–1661.

Brainard, D. H., & Wandell, B. A. (1992). Asymmetric color matching—How color appearance depends on the illuminant. *Journal of the Optical Society of America A: -Optics, Image Science, & Vision, 9*(9), 1433–1448.

Brown, R. O., & MacLeod, D. I. A. (1997). Color appearance depends on the variance of surround colors. *Current Biology, 7*(11), 844–849.

Buchsbaum, G. (1980). A spatial processor model for object color perception. *Journal of the Franklin Institute: Engineering & Applied Mathematics, 310*(1), 1–26.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Burnham, R. W., Evans, R. M., & Newhall, S. M. (1957). Prediction of color appearance with differ-

ent adaptation illuminations. *Journal of the Optical Society of America*, *47*(1), 35–42.

Chichilnisky, E. J., & Wandell, B. A. (1999). Trichromatic opponent color classification. *Vision Research*, *39*(20), 3444–3458.

Delahunt, P. B., & Brainard, D. H. (2004). Color constancy under changes in reflected illumination. *Journal of Vision*, *4*(9):8, 764–778, http://www.journalofvision.org/content/4/9/8, doi:10.1167/4.9.8. [PubMed] [Article]

Farnsworth, D. (1957). The Farnsworth-Munsell 100 hue test for the examination of colour discrimination. Baltimore, MD: Munsell Color Co. Inc.

Fasano, G., & Franceschini, A. (1987). A multidimensional version of the Kolmogorov-Smirnov test. *Monthly Notices of the Royal Astronomical Society*, *225*(1), 155–170.

Foster, D. H. (2003). Does colour constancy exist? *Trends in Cognitive Sciences*, *7*(10), 439–443.

Foster, D. H. (2011). Color constancy. *Vision Research*, *51*(7), 674–700.

Foster, D. H., & Nascimento, S. M. (1994). Relational colour constancy from invariant cone-excitation ratios. *Proceedings of the Royal Society of London B*, *257*(1349), 115–121.

Golz, J., & MacLeod, D. I. A. (2002). Influence of scene statistics on colour constancy. *Nature*, *415*(6872), 637–640.

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, *9*(11), 1367–1368.

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]

Heider, E. R. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, *93*(1), 10–20.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.

Hordley, S. D. (2006). Scene illuminant estimation: Past, present, and future. *Color Research & Application*, *31*(4), 303–314.

Ishihara, S. (1972). *Tests for colour-blindness*. Tokyo, Kyoto: Kannehara Shuppan Co., Ltd.

Jameson, D., & Hurvich, L. M. (1955). Some quantitative aspects of an opponent-colors theory. I. Chromatic responses and spectral saturation. *Journal of the Optical Society of America*, *45*(7), 546–552.

Jameson, D., & Hurvich, L. M. (1964). Theory of brightness and color contrast in human vision. *Vision Research*, *4*(1–2), 135–154.

Jin, E. W., & Shevell, S. K. (1996). Color memory and color constancy. *Journal of the Optical Society of America A*, *13*(10), 1981–1991.

Kay, P., Siok, W. T., Wang, W. S. Y., Chan, A. H. D., Chen, L., Luke, K. K., et al. (2009). Language regions of brain are operative in color perception. *Proceedings of the National Academy of Sciences of the USA*, *106*(20), 8140–8145.

Kulikowski, J. J., & Vaitkevicius, H. (1997). Colour constancy as a function of hue. *Acta Psychologica (Amst)*, *97*(1), 25–35.

Land, E. H. (1964). Retinex. *American Scientist*, *52*, 247–264.

Land, E. H., & McCann, J. J. (1971). Lightness and retinex theory. *Journal of the Optical Society of America*, *61*(1), 1–11.

Ling, Y., & Hurlbert, A. (2008). Role of color memory in successive color constancy. *Journal of the Optical Society of America A: Optics, Image Science, & Vision*, *25*(6), 1215–1226.

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

Malo, J., & Luque, M. J. (2002). COLORLAB: A color processing toolbox for Matlab. Internet site: http://www.uv.es/vista/vistavalencia/software.html

Maloney, L. T., & Wandell, B. A. (1986). Color constancy: A method for recovering surface spectral reflectance. *Journal of the Optical Society of America A: Optics, Image Science, & Vision*, *3*(1), 29–33.

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*(19), 3067–3078.

Nemes, V. A., McKeefry, D. J., & Parry, N. R. A. (2010). A behavioural investigation of human visual short term memory for colour. *Ophthalmic & Physiological Optics*, *30*(5), 594–601.

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

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

www.journalofvision.org/content/10/9/16, doi:10.1167/10.9.16. [PubMed] [Article]

Otazu, X., Parraga, C. A., & Vanrell, M. (2010). Towards a unified model for chromatic induction. *Journal of Vision*, *10*(12):5, 1–24, http://www.journalofvision.org/content/10/12/5, doi:10.1167/10.12.5. [PubMed] [Article]

Peacock, J. A. (1983). Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, *202*(2), 615–627.

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

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]

Shevell, S. K., & Wei, J. P. (2000). A central mechanism of chromatic contrast. *Vision Research*, *40*(23), 3173–3180.

Smith, V. C., & Pokorny, J. (1975). Spectral sensitivity of the foveal cone photopigments between 400 and 500 nm. *Vision Research*, *15*, 161–171.

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]

Smithson, H. E. (2005). Sensory, computational, and cognitive components of human colour constancy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1458), 1329–1346.

Speigle, J. M., & Brainard, D. H. (1996). Is color constancy task independent? In *Proceedings of the 4th Information Science and Technology/Society for Information Display Color Imaging Conference: Color Science, Systems, and Applications*, pp. 167–172. Springfield, VA: Society for Imaging Science and Technology.

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: -Optics, Image Science, & Vision, 16*(10), 2370–2376.

Troost, J. M., & de Weert, C. M. (1991). Naming versus matching in color constancy. *Perception & Psychophysics*, *50*(6), 591–602.

Von Kries, J. (1970). Chromatic adaptation. In D. L. MacAdam (Ed.) *Sources of Color Vision*, pp. 145–148. Cambridge, MA: MIT Press. (Originally published in Festschrift der Albrecht-Ludwigs-Universitat [1902]).

Webster, M. A., & Kay, P. (2007). Individual and population differences in focal colors. In R. E. MacLaury, G. V. Paramei, & D. Dedrick (Eds.), *Anthropology of color: Interdisciplinary multilevel modeling* (pp. 29–54). Philadelphia: J. Benjamins Publishing Co.

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

Yang, J. N., & Shevell, S. K. (2002). Stereo disparity improves color constancy. *Vision Research*, *42*(16), 1979–1989.

Appendix A

Given that our experiments were conducted over a few weeks, we tested whether the uncertainty introduced by longer-term memory was significantly larger than the uncertainty present in a typical 25-min session. We did this by repeating the same measures over different days using two experienced subjects. They were required to select four SRs (green, purple, orange, and gray) and to reproduce the same colors seven days later. To collect more data, the selection of SRs was repeated forty times for each color. Figure A.1 shows the variability of our measures for these control sessions: The small darker points correspond to results for the first session and the small lighter points to the second session. Squares and triangles represent the corresponding averages. The lightness variability results followed a similar trend and were omitted from the plots for clarity's sake. To determine if both distributions of points are the same, we computed the statistic D, the maximum difference of the integrated probabilities of the two distributions, developed by Fasano and Franceschini (1987) and others (Peacock, 1983). Our results showed that, predictably there were memory effects in all cases except two. However, D was comparatively small, i.e., the mean's difference between the light and dark points was always smaller than the standard deviation (itself about 1 ΔE*) of either the light or the dark point distributions.
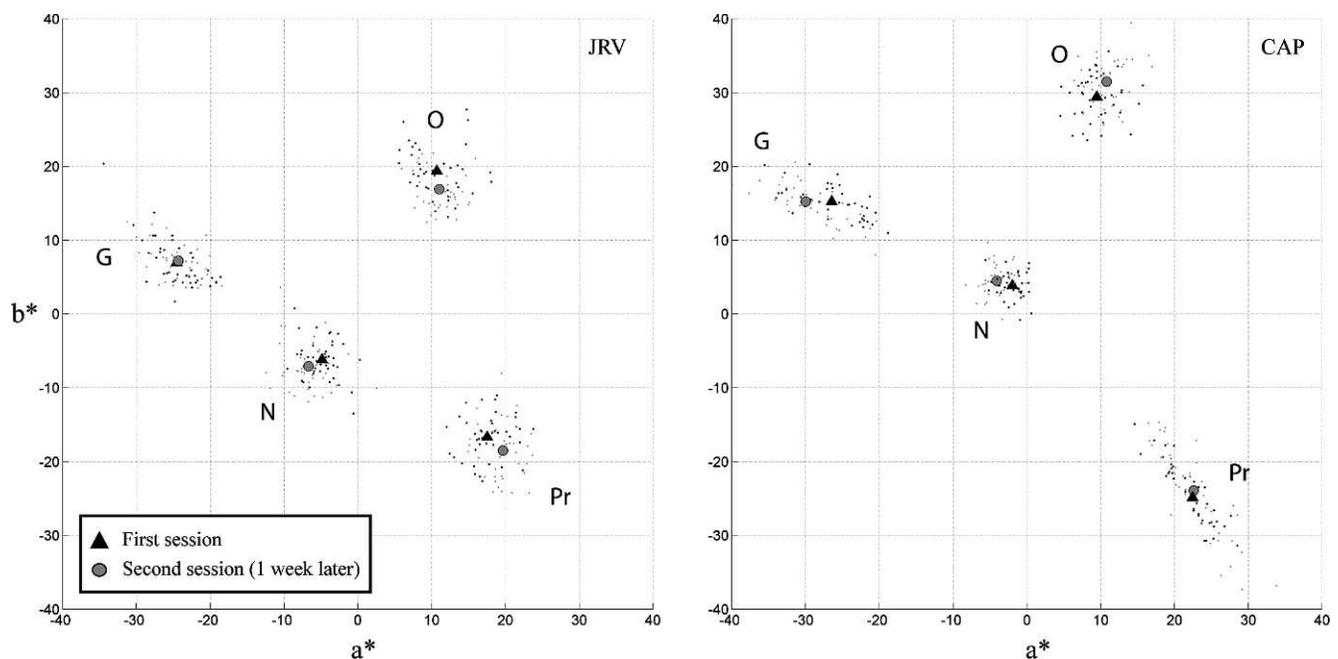
Figure A.1. Results of the long-term memory control experiment for two subjects. Four categories were tested (40 trials each). Dark and light dots were measured with a seven-day time difference. Averages are represented by triangles (first session) and squares (second session). The results clustered near the origin, are equivalent to those of a typical achromatic setting experiment.

Appendix B

The *Akaike Information Criterion* (AIC) method is based on Information Theory, and it is widely used for model selection, i.e., given several candidate models the method selects the model, which minimizes the loss of information when approximating the reality. In order to test our four models we used the AIC version adapted to small sets of samples (AIC_c) and the residual sum of squares (RSS) as detailed in Equation 9, where n corresponds to the number of data points and k to the number of variables plus the error term (Burnham & Anderson, 2002).

$$AIC_c = nln\left(\frac{RSS}{n}\right) + 2k + \frac{2k(k+1)}{n-k-1} \quad (9)$$

Notice that AIC_c formulae depends exclusively on the dimensions of the multivariate system resulting from Equation 5, because this approach does not reflect the number of free parameters existent in our tested models we rearranged the system into an equivalent univariate system. In order to apply the AIC_c we assumed that our prediction errors followed a normal distribution.

The model with the lowest AIC value is the best model among all models specified. However, AIC values become interesting when compared to the AIC value of a series of models. Two measures associated with AIC can be used to compare models: (a) the difference between the model with the lowest AIC and the rest ($\Delta_i = AIC_i - \min(AIC_i)$) and (b) the *Akaike weights*, which quantify the plausibility of each model as being the best ($w_i = \exp(-0.5\Delta_i)/\sum_{r=1}^{R}\exp(-0.5\Delta_r)$). As a rule of the thumb, a $\Delta_i < 2$ suggests substantial evidence for the model, values between 3 and 7 indicate that the model has considerably less support, whereas $\Delta_i > 10$ indicate that the model is very unlikely (Burnham & Anderson, 2002).

Table B.1 contains the values of the RSS, AIC_c, Δ_i and w_i when applied to our data according to the model, number of fitting points and illumination used. Notice that the reported RSS values do not correspond to the minimization ones in Figure 6; this is because we took as RSS value the accumulative error of the fitting points that participated in the minimization process only. In practice, RSS values were not obtained by linear regression but from the minimization process described in Equation 6; however, the target value of the minimization is equivalent. Also the RSS values used in Table B.1 resulted from the average over all subjects and backgrounds.

Model	3n	k	Greenish				Yellowish			
			RSS	AIC_c	Δ_i	w_i	*RSS*	AIC_c	Δ_i	w_i
D	15	4	293.18	56.59	0	0.96	535.1	65.62	0	0.99
DT	15	7	134.79	62.93	6.34	0.04	298.2	74.84	9.23	0.01
L	15	10	109.74	104.85	48.26	0	218.6	115.18	49.57	0
A	15	13	40.96	405.01	348.48	0	81.1	415.31	349.69	0
D	18	4	367.05	65.35	0	0.77	671.6	76.22	0	0.95
DT	18	7	187.90	67.42	2.07	0.26	425.9	82.15	5.92	0.05
L	18	10	164.36	91.24	25.89	0	328.4	103.70	27.47	0
A	18	13	82.27	144.36	79.00	0	170.5	157.47	81.25	0
D	21	4	440.95	74.43	0.44	0.45	808.3	87.16	0	0.88
DT	21	7	242.55	74.00	0	0.55	547.2	91.08	3.92	0.12
L	21	10	218.61	91.20	17.20	0	434.7	105.63	18.47	0
A	21	13	123.84	115.26	41.27	0	270.1	131.63	44.48	0
D	24	4	514.87	83.69	2.27	0.24	945	98.26	0	0.83
DT	24	7	297.47	81.41	0	0.76	685	101.45	3.19	0.17
L	24	10	273.25	95.30	13.88	0	550	112.07	13.81	0
A	24	13	169.47	109.31	27.90	0	381.9	128.81	30.55	0
D	27	4	588.79	93.04	3.72	0.13	1081.8	109.46	0	0.70
DT	27	7	353.17	89.31	0	0.86	794	111.19	1.72	0.29
L	27	10	327.65	101.14	11.83	0	617	118.23	8.77	0.01
A	27	13	226.17	111.39	22.07	0	495.3	132.55	23.10	0

Table B.1. The Akaike Information Criterion applied to our data. *Notes*: Each row corresponds to the model case considered, and the columns correspond to the number of fitting points used, the number of free parameters in each model, the RSS, and the Akaike results: AIC_c, Δ_i, and w_i. Note that the multivariate system was rearranged into an equivalent univariate system, therefore the 3n factor in the second column. See details on how these values were computed in the main text.

Δ_i and w_i values in Table 6 indicate that the Diagonal and Diagonal plus Translation models are the ones that best model the data, and indicate that the Linear and Affine models significantly over-fit the data. The small differences in Δ_i between D and DT are not conclusive about which is the best model; however, there is a clear tendency as we add more fitting points; the DT model becomes better than D. From one to four fitting points the AIC indicates that the best model is the D, DT, L, and A as expected due to the coincidence between the number of fitting points and the free parameters.

Appendix C

	New observers (2)		Original observers (10)	
Index/Illuminant	Greenish	Yellowish	Greenish	Yellowish
$\overline{BR}$	0.70	0.80	0.62	0.61
CCI	0.74	0.87	0.76	0.75
SCI	1.15	0.84	1.03	0.85

Table C.1. Color constancy indices of two observers from the new control experiment (left columns) and the rest of observers (right columns; see Table 5).

In our experiments, observers performed chromatic settings on a small percentage of patches over illuminated backgrounds Type I (which already contained selected representatives). This is potentially very problematic, since observers might learn the correct settings when they see the correct answer in the Type I backgrounds. However, this is very unlikely to happen in practice, given that observers see SRs illuminated by colored light and presented on a variegated form (not on a Type 0 background as they originally saw them). Moreover, all observers (except the authors) did not know whether they were related in any way to their reference session selection.

To address this potential problem we performed a control experiment with two new subjects who previously had only seen Type II backgrounds. The experiment consisted of selecting SRs with Type II background and all three illuminants. Figure C.1 shows these results, where chromatic settings (averaged for the two observers) are shown as filled circles and the

average chromatic settings for the rest of observers (see Figure 6) are shown as squares. Panels A, B, and C of Figure C.1 show results for the three illuminants (D65, greenish, and yellowish respectively). To facilitate the comparison, panels D, E, and F show the same results replotted using the gray SR as a reference white point.

Table C.1 shows the values (averaged over these two new subjects and backgrounds) for the three indices discussed in the manuscript. If having seen chromatic settings in reference sessions confers any advantage, the two new subjects should consistently have lower index values than the rest. The results in Table C.1 do not support this.
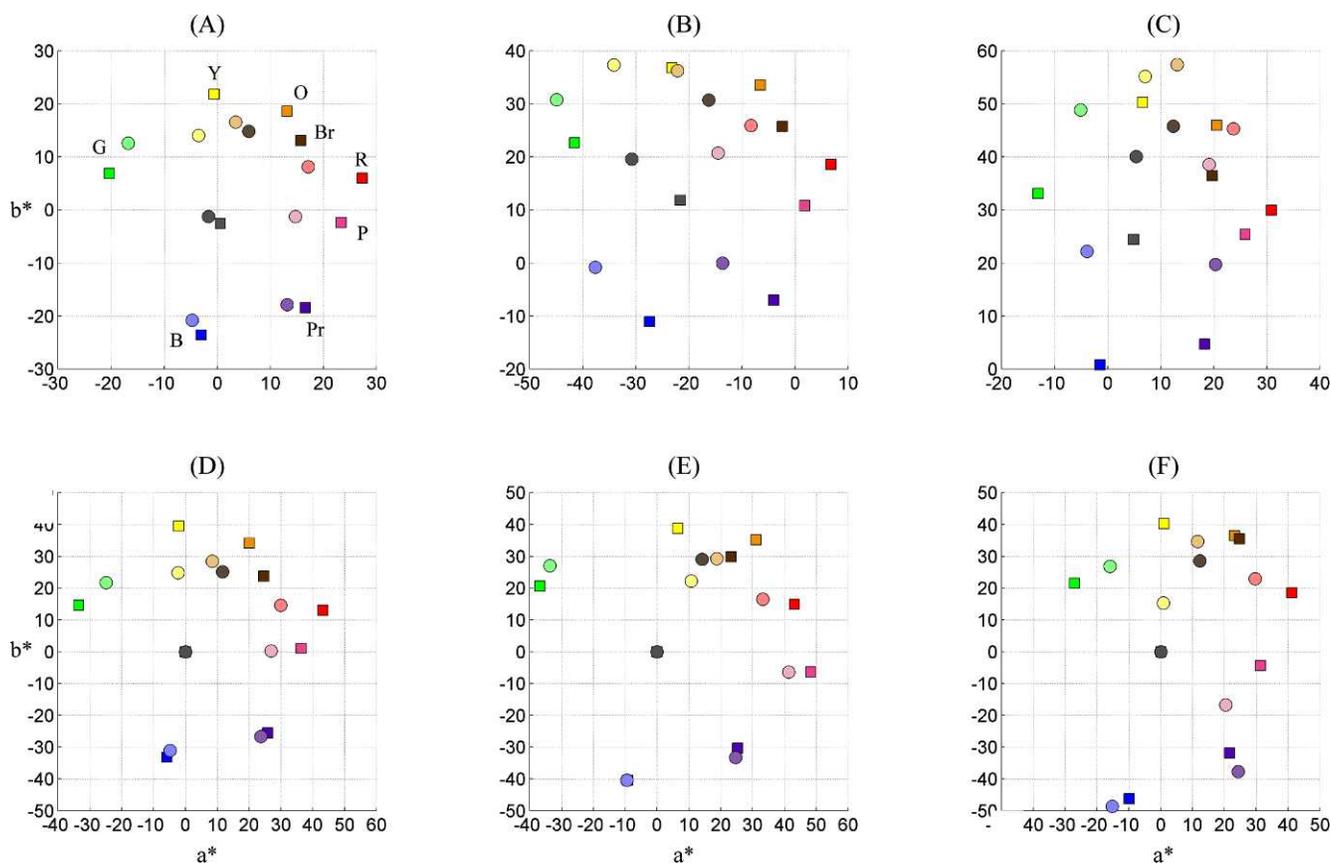
Figure C.1. Chromatic settings of two observers from the control experiment. This experiment tested the effect of using the chromatic settings as background colors, i.e., none of them saw Background Type I. Squares represent the results for the two new observers and circles represent the average for the original ten observers (see Figure 6). Bottom panels show the same data as in top panels, using the "gray" SR as a reference white point. As before, markers are color-coded according to color categories.