Color and gloss constancy under diverse lighting environments

Takuma Morimoto
Justus Liebig University Giessen, Giessen, Germany
Department of Experimental Psychology, University of Oxford, Oxford, UK

Arash Akbarinia
Justus Liebig University Giessen, Giessen, Germany

Katherine Storrs
Justus Liebig University Giessen, Giessen, Germany
School of Psychology, University of Auckland, New Zealand

Jacob R. Cheeseman
Justus Liebig University Giessen, Giessen, Germany
Center for Mind, Brain and Behavior (CMBB), Universities of Marburg, Giessen and Darmstadt, Germany

Hannah E. Smithson
Department of Experimental Psychology, University of Oxford, Oxford, UK

Karl R. Gegenfurtner
Justus Liebig University Giessen, Giessen, Germany

Roland W. Fleming
Justus Liebig University Giessen, Giessen, Germany
Center for Mind, Brain and Behavior (CMBB), Universities of Marburg, Giessen and Darmstadt, Germany

When we look at an object, we simultaneously see how glossy or matte it is, how light or dark, and what color. Yet, at each point on the object's surface, both diffuse and specular reflections are mixed in different proportions, resulting in substantial spatial chromatic and luminance variations. To further complicate matters, this pattern changes radically when the object is viewed under different lighting conditions. The purpose of this study was to simultaneously measure our ability to judge color and gloss using an image set capturing diverse object and illuminant properties. Participants adjusted the hue, lightness, chroma, and specular reflectance of a reference object so that it appeared to be made of the same material as a test object. Critically, the two objects were presented under different lighting environments. We found that hue matches were highly accurate, except for under a chromatically atypical illuminant. Chroma and lightness constancy were generally poor, but these failures correlated well with simple image statistics. Gloss constancy was particularly poor, and these failures were only partially explained by reflection contrast. Importantly, across all measures, participants were highly consistent with one another in their deviations from constancy. Although color and gloss constancy hold well in simple conditions, the variety of lighting and shape in the real world presents significant challenges to our visual system's ability to judge intrinsic material properties.

Introduction

In our everyday lives, we often identify and interact with objects across major changes in lighting—for example, returning at sunset to a car we parked in the morning or carrying a familiar coffee mug from the kitchen to the balcony. In such circumstances, we are not usually struck by the impression that the color of the car's paint or the gloss of the mug's glaze have changed. Yet judging the color and gloss of a surface—especially across different lighting conditions—poses hard computational challenges. First, these perceptual quantities result from two physically distinct aspects of the surface's reflectance,

Citation: Morimoto, T., Akbarinia, A., Storrs, K., Cheeseman, J. R., Smithson, H. E., Gegenfurtner, K. R., & Fleming, R. W. (2023). Color and gloss constancy under diverse lighting environments. *Journal of Vision*, *23*(7):8, 1–25, https://doi.org/10.1167/jov.23.7.8.

https://doi.org/10.1167/jov.23.7.8 Received December 9, 2022; published July 11, 2023 ISSN 1534-7362 Copyright 2023 The Authors

which are confounded in the retinal image. Surface color typically depends primarily on diffuse reflection in which the incident illumination is modified through interaction with pigments, whereas gloss stems from specular reflection, which is a direct reflection of the illuminant. At each point on the object's surface, diffuse and specular reflections are mixed in different proportions, resulting in substantial spatial chromatic and luminance variations. In the demonstration in Figure 1a, pixel colors vary dramatically among three selected points on the surface of the object, showing that there is no simple mapping between the cone excitations at a single pixel and the color of the object. Furthermore, when we created this image (using computer graphics rendering techniques), we placed an object under a specific lighting environment, applied specific diffuse and specular reflectance properties to the object, and fixed the camera at a specific viewpoint (Figure 1a, top right). Changing any of these underlying scene parameters would alter the color distribution across the same object's surface in the resultant image. As shown in Figure 1b, pixel colors of the same three surface points change substantially when the same object is viewed in the same pose, but in different lighting environments. The reader may also have the impression that the color and glossiness of the object appear to be somewhat different, even though the material properties are identical across the three objects.

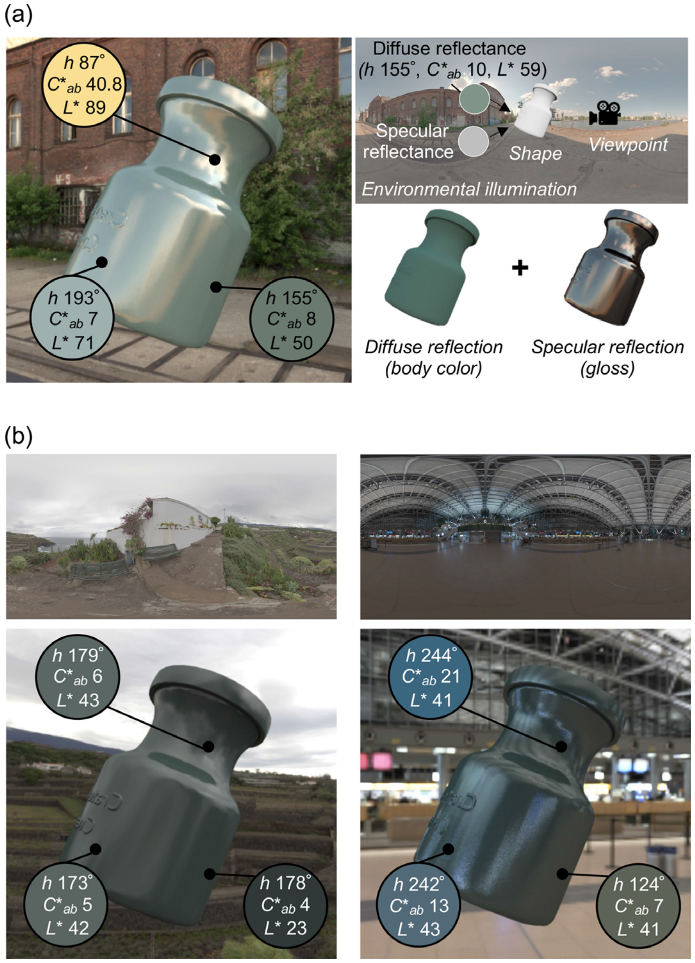

Figure 1. Graphical illustration of challenges to judging the color and gloss of a three-dimensional object placed under a complex lighting environment. (a) As shown to the upper right, we placed an object under an environmental illumination, applied a diffuse reflectance, which has a fixed hue, chroma, and lightness (denoted as h, C^*_{ab} and L^*, respectively), and a specular reflectance, set a view-point and rendered the object image using computer graphics techniques. The rendered image is shown on the left side. The hue, chroma, and lightness vary largely across the three selected pixels at different regions on the object's surface. (b) Effects of lighting environments (overcast environment on the left and indoor environment on the right) on the color variation at the same three locations. Note that the objects in (a) and (b) have identical material properties though their appearance may differ across lighting environments.

Color constancy has been a core domain of human color vision research (Foster, 2011; Hurlbert, 2007; Smithson, 2005). Many early studies used simplistic visual stimuli that were flat, matte, and uniformly illuminated, but more recent studies have begun to use more complex stimuli that better represent natural visual environments (Brainard & Maloney, 2004; Mizokami, 2019; Witzel & Gegenfurtner, 2018). One common research question is whether the use of real three-dimensional objects increases the degree of color constancy, compared to flat and uniformly colored surfaces or images presented on a computer monitor (de Almeida, Fiadeiro & Nascimento, 2010; Hedrich, Bloj & Ruppertsberg 2009; Morimoto, Mizokami, Yaguchi, & Buck, 2017). Color and lightness constancy in the presence of specular reflection have also been well-studied, and several studies have found that the presence of specular reflections can improve color constancy (Hurlbert, Cumming, & Parker, 1991; Lee & Smithson, 2016; Granzier, Vergne & Gegenfurtner, 2014; Nagai, Kaneko, Kawashima, & Yamauchi, 2017; Snyder, Doerschner & Maloney, 2005; Yang & Maloney, 2001; Yang & Shevell, 2003), whereas other works found the effect only in limited conditions (Xiao, Hurst, MacIntyre, Brainard, 2012; Wedge-Roberts et al., 2020).

The effect of object and lighting properties on perceived gloss has also been the subject of several studies, although the term "constancy" is not always explicitly stated. Many factors that can alter perceived gloss have been identified: for example, object motion (Doerschner et al., 2011), body color (Wendt, Faul, Ekroll, & Mausfeld, 2010), surface curvature (Ho, Landy, & Maloney, 2008), and lighting environment statistics (Adams, Kucukoglu, Landy, & Mantiuk, 2018; Fleming, Dror & Adelson, 2003; Obein, Knoblauch, & Viéot, 2004; Pont & te Pas, 2006). Furthermore, systematic perturbation of lighting and material properties have been used to quantify their influences on perceived gloss level (Zhang, Ridder, & Pont, 2018; Zhang, de Ridder, Barla, & Pont, 2019, 2020). Motoyoshi and Matoba (2012) used scenes containing a statue along with other objects. They showed failures of gloss constancy and that surrounding context information had virtually no effect on matching results, implying that participants do not use contextual information to discount the influence of illumination.

In addition to these constancy perspectives, specific visual computations that may underlie gloss perception have been proposed in recent decades (Chadwick & Kentridge, 2015; Landy, 2007). One primary discussion point is whether the visual system makes use of summary statistics extracted from a given image, rather than reconstructing the entire optical input (Fleming, 2014, 2017; Nishida, 2019). Candidate cues to gloss include skewness of the luminance histogram (Anderson & Kim, 2009; Kim & Anderson, 2010; Motoyoshi, Nishida, Sharan, & Adelson, 2007; Sharan, Motoyoshi, Nishida, & Adelson, 2008), its standard deviation (Wiebel, Toscani & Gegenfurtner, 2015), the magnitude of luminance gradients in an image (Sawayama & Nishida, 2018), and more complex image metrics computed from specular reflection patterns (Marlow & Anderson, 2013; Marlow, Kim, & Anderson 2012). Perceived gloss is nonlinearly related to underlying physical quantities such as the proportion of light reflected in a specular manner, and the discriminability of gloss as a function of specular reflectance has been recently investigated (Cheeseman, Ferwerda, Maile, & Fleming, 2021).

Despite these extensive investigations into color constancy and gloss constancy as separate domains, very few studies have simultaneously measured color and gloss constancy (lightness and gloss [Hansmann-Roth & Mamassian, 2017; Olkkonen & Brainard, 2010], gloss and hue variation between green and blue [Brainard, Cottaris & Radonjić, 2018; Radonjić, Cottaris, & Brainard, 2018, 2019]). No study has measured all dimensions of perceived color (hue, chroma, and lightness) and gloss at the same time. Yet, in daily life, when we look at an object, both percepts naturally occur together, and we do not judge each separately unless asked to do so explicitly. The purpose of this study was to directly measure our ability to judge color and gloss using a synthetic image set produced using physically based ray-tracing techniques from computer graphics that captures large variations of object properties and illuminant properties. For this, we used a well-established asymmetric matching task in which two images were presented side by side on a computer screen. Participants were asked to adjust the color and gloss of a right reference object so that it appears to be made of the same material as a left test object. The critical manipulation is that the two objects were presented under different lighting environments; thus, participants needed to take this difference into account to achieve accurate matching of physical reflectance parameters. This methodology has been extensively used in perceptual constancy studies and shown to be an effective constancy task (Arend & Reeves 1986; Brainard & Wandell, 1992). For instance, Olkkonen and Brainard (2010) jointly measured perceived lightness and gloss of smooth spheres placed under real-world lighting environments. They showed that participants accurately judged lightness, but perceived gloss varied substantially across lighting environments. There are a few other studies using the asymmetric matching task that did not consider the influence of illuminant changes, but that are nonetheless relevant to the present study. Xiao and Brainard (2008) showed that humans do not simply use the mean color across the whole object to determine the overall color impression of a three-dimensional glossy object. Strong interactions between specular reflection, chroma and lightness were found (Honson et al. 2020; Isherwood et al., 2021). Color matching using different types of materials (papers, sponges, wool, candles, and porcelain) revealed that hue perception is highly stable, while chroma and lightness were influenced by material types (Giesel & Gegenfurtner 2010).

This study built on these past efforts in two ways: i) participants were asked to judge three color dimensions (hue, lightness, and chroma) and gloss at the same time, and ii) we used a substantially greater variety of lighting environments and object shapes to capture the complex behavior of constancy mechanisms (including successes and errors) in a large stimulus space. Two psychophysical experiments were performed. The first experiment measured perceived color and gloss from single objects, under different lighting environments, where each object was assigned a unique color and gloss level randomly. The second experiment followed up the failure of gloss constancy observed in the first experiment by systematically exploring the effects of object shape and lighting environment on perceived gloss.

General methods

Apparatus

Both experiments were computer-controlled and all images were displayed on a 24-inch LCD monitor (ColorEdge CG2420, 1920 × 1200 pixels, frame rate 60 Hz; EIZO, Ishikawa, Japan) in 10 bits per color channel (red, green, and blue). We performed gamma correction and spectral calibration using a spectroradiometer (CS-2000; KONICA MINOLTA, Inc., Tokyo, Japan). Experimental code was written in MATLAB using custom-built functions as well as functions provided in PsychToolbox-3 (Brainard, 1997).

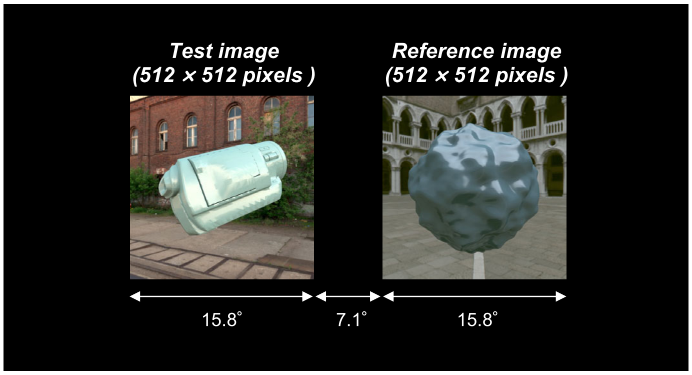

Figure 2. Stimulus configuration for asymmetric matching in this study. The task of participants was to adjust an underlying diffuse reflectance to change the color and an underlying specular reflectance to change the gloss of the reference object presented in the right reference image until it appear to be made of the same material as the test object presented in the left test image. Test images changed from one trial to another, but the reference image stayed the same throughout the experiment. White text and arrows were not presented during the actual experiment.

Participants

Ten naive participants were recruited for Experiments 1 and 2. The ratio of female to male was 2.33 for Experiment 1 and 1.50 for Experiment 2. The age of participants ranged from 21 to 32 years, with a mean of 25.6 ± 2.83 years for Experiment 1 and 24.0 ± 3.33 years for Experiment 2. Four participants completed both experiments. The experiments were approved by a local ethics committee at Justus Liebig University Giessen in accordance with the Helsinki Declaration (sixth revision, 2008). All participants had normal or corrected-to-normal visual acuity (self-report). Before the experiments, all participants were screened for normal color vision using Ishihara pseudoisochromatic testing plates (Ishihara, 1973). All participants were undergraduate students at Justus Liebig University Giessen, Germany, and were paid for their time.

Task

We used an asymmetric matching task. As shown in Figure 2, a test image and a reference image were presented side by side, separated by 7.1° of visual angle. Participants were asked to adjust the color and gloss of the right reference object until it appeared to be made of the same material as the left test object. The adjustment of color and gloss were done in the object space rather than the image space. In other words, for the color setting, participants changed an underlying diffuse reflectance of the reference object by adjusting hue, lightness (L^*) and chroma (C^*_{ab}) (as defined in the $L^*a^*b^*$ color space, for the reflectance under equal energy white) which updated the weights of reflectance basis functions (described elsewhere in this article) to produce a composite reflectance with the desired color properties. For the gloss setting, participants changed an underlying specular reflectance by adjusting a parameter c, a perceptually linear gloss scale developed by Pellacini, Ferwerda, and Greenberg (2000) (Pellacini's c hereafter). We used these approximately perceptually linear scales for color and gloss adjustment because we predicted that participants might find it easier if increasing or decreasing a single step would have approximately equal perceptual effect at any point on each parameter range. Participants used eight buttons, each corresponding with increasing or decreasing values of one of the four parameters. A beep was provided when the value reached the limit of the prepared range for lightness, chroma, and Pellacini's c. Hue is a circular variable, and thus there was no range limit. Participants binocularly viewed stimuli presented on a flat screen. During the matching task, participants were allowed to move their eyes freely between the test image and the reference image. There was no time limit for each trial. In this study, the right reference object served as a scale to quantify participants' subjective experiences, and thus the shape and the lighting environment in the reference image stayed the same during the whole experiment, and only the left test image changed from one trial to the next. A key feature of this task was that the two objects were placed under different lighting environments. Thus, participants needed to discount the effect of lighting on appearance to make accurate color and gloss matches. Pixel colors have a complex relationship with the physical parameters applied to the test object because they are influenced by the lighting environment, as shown in Figure 1. However, a perfectly color- and gloss-constant observer should be able to estimate physical parameters from the test image regardless of the lighting environment and assign the same parameters to the reference image. Details of test images and reference images will be explained elsewhere in this article.

Experimental stimuli

Rendering

All images used in this study were generated using the physically based renderer Mitsuba (Jakob, 2010). A single image contained one floating object shape, to which we applied a diffuse reflectance and a specular reflectance using the balanced Ward reflectance model (Ward, 1992) that conserves energy at grazing angles (Geisler-Moroder & Dür, 2010). For the illumination, we used image-based lighting (as shown in Figure 1), which we call environmental illumination in this study, where each pixel in the environmental illumination map conveys information about a light ray that reaches a single point in the scene from a particular direction. The map thereby depicts light coming from every possible direction in the environment. Environmental illuminations used in this study were all originally RGB images, and Mitsuba was used to convert these to hyperspectral images during the rendering process (Smits, 1999). To maximize the accuracy of the rendering process, all images were spectrally rendered from 400 to 720 nm in 10-nm steps (31 spectral channels). All images were rendered in 512 × 512 pixel resolution. Rendered hyperspectral images were converted to the XYZ color space to calculate image statistics, and to the calibrated RGB color space, such that displayed images reproduced the XYZ values from the rendered images.

Reference images

Measuring perceived color and gloss using the asymmetric matching task required a reference scene in which participants can adjust the physical reflectance parameters of the reference object in a continuous fashion. To construct the reference scene, we selected the i) environmental illumination, ii) shape of the reference object, and iii) diffuse and specular reflectances as follows. In selecting the reference environment, we wanted an environment in which chromatic variation is low and a sufficient amount of light hits the objects embedded in this environment directly, producing highly visible specular reflections, which we expected could help participants to judge easily both the color and gloss of the reference object. We also wanted to choose an environment in which the reference object appears to be highly glossy, because otherwise participants may not find a satisfactory match with a test object, even at the highest gloss level allowed for the reference object. To establish this circumstance, we plotted diagnostic metrics, for publicly available maps, as shown in Figure 3. Figure 3a shows the selected environmental illumination map (Uffizi Gallery, Italy) downloaded from a publicly available database (Debevec, 1998; https://vgl.ict.usc.edu/Data/HighResProbes/; accessed March 15, 2022) along with statistical characterization of the illumination. The top right plot shows a luminance histogram of all pixels together with some basic statistics: mean, standard deviation, skewness, kurtosis, and Xia's diffuseness metric (Xia, Pont, & Heynderickx, 2017), which quantifies how much the lighting environment is directionally uniform from 0 (point light source) to 1 (fully uniform). This lighting environment has low diffuseness because the upper and lower hemifields have largely different directional lighting patterns. The bottom left plot shows 10% pixel distribution on a^*b^* chromatic plane where equal energy white is set to the origin. The bottom right shows a power spectrum, analyzed by spherical harmonic decomposition, with a negative slope of the regression line about –2 that was shown to be typical for outdoor environmental illuminations (Dror, Wikksky, & Adelson, 2004).

For the reference shape, we chose a simple, bumpy sphere whose surface roughness was fixed at 0.05 as defined in Mitsuba. We chose this shape as a standard shape again because some degree of bumpiness increases the spatial contrast of specular reflection between concave and convex regions, helping participants to detect where the specular reflections are, and high spatial-frequency geometries have been associated with better material recognition accuracy (Lagunas, Serrano, Gutierrez, & Masia, 2021).

For a continuous adjustment of color and gloss, we needed to systematically control the underlying physical reflectances, separately for diffuse reflectance and specular reflectance. Our approach was to separately render diffuse and specular images which were linearly summed to produce a colored glossy object, which provided a good enough approximation for the geometry considered here.

We first generated a specular image that has only a specular reflection component by setting the specular reflectance to 1.0 and diffuse reflectance to zero across all wavelengths. To modulate the gloss level of test images, this specular image was multiplicatively scaled up and down by a single factor. As already elsewhere in this article, participants adjusted the gloss level by changing Pellacini's c, and this parameter was directly converted to specularity p_s, a parameter defined in the Ward reflectance model. The scalar quantity p_s was used to scale the specular image at each wavelength. We sampled Pellacini's c from 0 (matte) to 0.149 (highly glossy), which corresponded with 0 and 0.0999 in specularity p_s, respectively. Pellacini's c captures the lightness-dependent nature of perceived gloss and, thus, requires a lightness value for the conversion between Ward's specularity and Pellacini's c. For this, we used a value of 50, which is roughly equivalent to the mid value of the prepared lightness range (detailed elsewhere in this article). Although fixing the assumed lightness level to the single value was a necessary choice to keep the lightness and Pellacini's c to be independent during the adjustment, this practice meant that the scale we prepared may not be perfectly linear especially when the actual lightness level is far from 50 (e.g., 28 or 74). However, in pilot work, we confirmed that Pellacini's c was still better than Ward's specularity in terms of perceptual linearity even considering this limitation.

Our approach to defining diffuse reflectances was to first specify colors in terms of hue, lightness, and chroma in $L^*a^*b^*$ color space and to find a spectral reflectance by combining basis reflectance functions extracted from the Munsell color system, that produces the desired hue, lightness and chroma

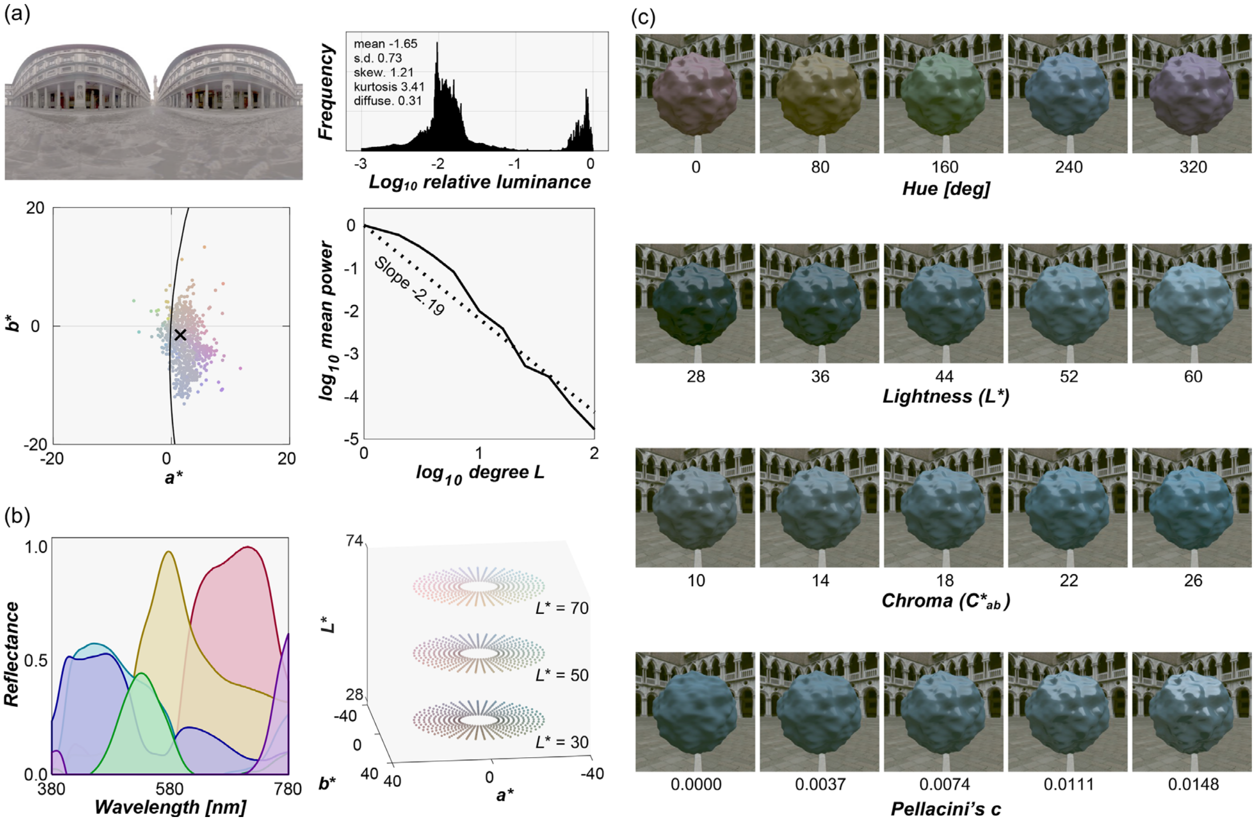

Figure 3. (a) Reference environmental illumination used for the reference image and its statistical characterization. The upper right plot shows the luminance histogram of all pixels, along with some image statistics (mean, standard deviation, skewness, kurtosis, and Xia's diffuseness metric). The lower left plot shows the a^*b^* chromatic distribution of 10% pixel colors that were sampled randomly. The lower right plot shows the power spectrum analyzed by spherical harmonic decomposition. (b) Six basis reflectance functions extracted from 1,269 matte Munsell color chips using non-negative matrix factorization. Subset of surface colors assigned to the object in the reference images at example lightness planes ($L^* = 30$, 50, and 70). There were in total 21,600 colors, allowing participants to explore the stimulus space using a method of adjustment. (c) Some example reference images, drawn from the four dimensional stimulus space of color and gloss adjustments.

when placed under an equal energy white light ($X = Y = Z = 100$).

We first applied non-negative matrix factorization to 1,269 matte Munsell color chips (https://sites.uef.fi/spectral/munsell-colors-matt-spectrofotometer-measured/, accessed March 15, 2022) to obtain six basis spectral reflectance functions as shown on the left in Figure 3b. Then, we sampled 21,600 colors in $L^*a^*b^*$ color space (subset is shown in the right plot of Figure 3b): 90 hue values from 0° to 356° in 4° steps; 24 lightness values from 28 to 74 in 2 steps; and 10 chroma values from 8 to 26 in 2 steps. Then, for each color, we searched for the optimal weights for six basis reflectances so that the weighted summation of the basis functions produced the desired L^*a^*b* values under equal energy white light ($X = Y = Z = 100$). For the optimization, we set a condition that the resultant diffuse reflectance value should be between zero and 0.90 at any wavelength to meet physical constraints, even when being combined with the specular reflectance (0.0999 at maximum). This process generated 21,600 spectral diffuse reflectances; using these, we rendered 21,600 matte reference images. We used 36 of these reflectances to generate test images in Experiment 1 as described elsewhere in this article. These diffuse images, along with the aforementioned single specular image, enabled continuous adjustment of the color and gloss of the objects in such a way as to achieve approximately uniform coverage of the perceptual space, at least as expected under reference viewing conditions. Figure 3c shows some example reference images.

When we rendered scenes in the Mitsuba renderer, the object was placed at the center of the scene and environmental illumination was applied to the scene. Then, the camera was set at the same height as the object and pointed directly at the object. Environmental illumination generally had high dynamic range, and had we selected the viewpoint (i.e., camera position) at random it would have often created an image in which the object region was too dark to see. Thus, we rendered a mirrored sphere from different viewpoints (0° to 330° in 30° steps) and picked the viewpoint at which the mean luminance over the sphere was the highest. We also made sure that there were no pixels with particularly intense lights in the surrounding non-object region of the image. Pixel values in the raw hyperspectral images from Mitsuba were defined arbitrarily because pixel values in the environmental illumination have no units. Thus, after converting the raw hyperspectral image to a linear monitor RGB image, we scaled the whole image by the 99th percentile pixel value across all reference images. We applied the same scaling value to all reference images to equate the light level across images. Because of these selection processes, no tone mapping was applied to any of the reference images. These scaled linear RGB images were gamma corrected and used for the experiment.

Test images

We first gathered environmental illuminations from multiple online databases (including Adams et al., 2016; Debevec, 1998; and https://hdrmaps.com/freebies/), from which we selected 12 lighting environments that cover a diverse variation of natural lighting. Figure 4 shows the selected lighting environments along with the luminance histogram of all pixels and 10% pixel chromatic distributions on a^*b^* plane for six outdoor scenes (Figure 4a) including sunny days (left four scenes) and cloudy days (right two scenes) and six indoor scenes (Figure 4b).

To expand the diversity of the selected lighting environments to unnatural domains, we applied two independent manipulations to each lighting environment. First, we rotated the a^*b^* chromatic distribution of the original lighting environment by +90° while keeping the L^* distribution the same, producing a gamut-rotated lighting environment. Chromaticities in natural lighting environments tend to cluster along a blue–yellow axis known as the CIE daylight locus (Hernández-Andrés, Romero, Nieves, & Lee, 2001; Judd et al., 1964), and this manipulation made the distribution run along an orthogonal red–green axis. If our visual system uses a prior about a typical illuminant to achieve perceptual constancy (Delahunt & Brainard, 2004; Pearce, Crichton, Mackiewicz, Finlayson, & Hurlbert, 2014; Weiss, Witzel, & Gegenfurtner, 2017), we may observe higher errors in participants' settings under these chromatically atypical environments. After the rotation, some pixels in the map went outside the chromatic gamut of the experimental monitor and were thus likely to produce out-of-gamut pixels in the rendered images. For those pixels, the chroma value was decreased until the colors came inside the chromatic gamut. When selecting the original lighting environments, we made sure that such pixels were always less than 5%. The second manipulation was to distort the directional structure of each lighting environment using a phase scrambling technique via spherical harmonic decomposition. This manipulation normally changes the color distribution of the image, but we kept the distribution of chromaticity and luminance by histogram matching the phase-scrambled environment maps to the originals. Again, if our visual system uses a prior about the directional structure, such as light from above (e.g., Morgenstern, Murray, & Harris, 2011; Ramachandran, 1988), we might observe poor perceptual constancy under these environments because intense lights could hit the object from every direction. The resultant 36 lighting environments (12 natural, 12 gamut-rotated, and 12 phase-scrambled) were used in Experiment 1. Twelve natural lighting environments were used in Experiment 2.

For Experiment 1, for each of the 36 lighting environments, we placed in the scene a randomly selected object from a dataset of three-dimensional meshes of everyday objects (purchased from Evermotion, https://evermotion.org/), and physical reflectance parameters (hue, chroma, lightness, and Pellacini's c), and selected the viewpoint in the same way we selected the viewpoint for the reference image. The viewpoint was shared between natural environments and gamut-rotated environments, but different camera angles were used for phase-scrambled environments (because the scene geometry was different). No tone mapping was applied to any of the test images. Figure 5 shows 36 example test images together with an example lighting environment and its a*b* chromatic distribution. Mean luminance over the object region was 22.6 ± 3.44 cd/m^2 (mean ± 1.0 standard error across 36 objects). For each test image, the underlying physical reflectance parameters were defined as ground-truth values in the analysis.

For Experiment 2, we generated 216 test images, a factorial combination of 12 natural lighting environments and 18 shapes sampled from shapes used in Experiment 1.

Procedure

Before the experiment, participants were given instructions both orally and by written text on the screen: "Your task is to adjust glossiness and color of the right 'reference' object in terms of specularity, lightness, chroma, and hue until the reference object appears to be made of the same material as the left 'target' object. Each parameter changes the appearance of the reference object as shown here." Regarding the final sentence, we presented some example reference images similar to Figure 3c to explain how changing each parameter influences the appearance of the

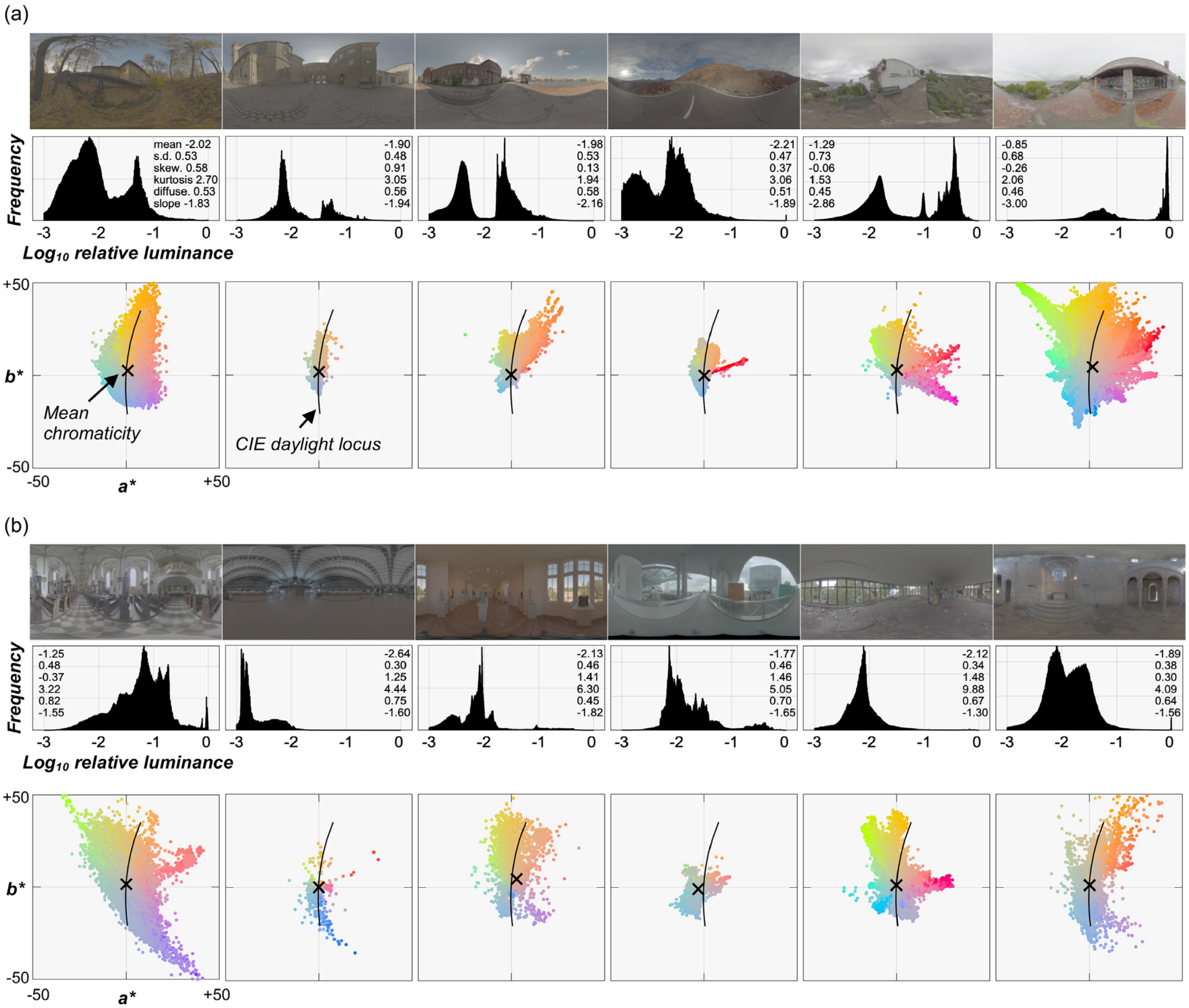

Figure 4. (a) Six outdoor lighting environments (left four environments are sunny days and the right two environments are overcast days) and (b) six indoor lighting environments, together with the luminance histogram of all pixels and the *a**b** chromatic distribution of 10% of pixel colors sampled at random. The numbers in the histograms are mean, standard deviation (SD), skewness (skew.), kurtosis, Xia's diffuseness metric (diffuse.), and slope of the power spectrum computed by spherical harmonic decomposition. Maximum luminance in each image was normalized to 1.0 (zero on a logarithmic scale) to allow the comparison across environments in this figure, but note that for actual test images each environment was scaled differently from this figure (detailed in main text). The intersection of horizontal and vertical thin gray lines denotes the white point ($X = Y = Z = 100$) of the color space. The black cross symbol depicts mean *a**b** chromaticity across the plotted 10% of pixels. The black solid line shows the CIE daylight locus.

reference object. Our instructions were phrased to encourage participants to judge the intrinsic properties of the material. We chose this criterion over matching in terms of 'appearance' (see discussion on the two criteria; e.g., Radonjic & Brainard, 2016; Reeves, Amano, & Foster, 2008), mainly because it better served the primary purpose of this study, which was to measure the extent to which participants can infer underlying material properties under diverse lighting environments. Given that this was an asymmetric matching task, with large spatial chromatic and luminance variation over the surfaces (Figure 1), it is questionable whether instructions focused on local appearance would have been as easy and consistent; depending on where observers look, the matching result would presumably change (Xiao & Brainard, 2008; Toscani, Valsecchi, & Gegenfurtner, 2013a, 2013b). Alternatively, if observers interpreted the appearance instructions to mean a spatial summary of the appearance across the surface, the resulting matches would likely closely resemble those resulting from our instructions. Three practice trials immediately followed the instructions, using example test images that were not used in the actual experiment. During the practice trials, participants were asked to

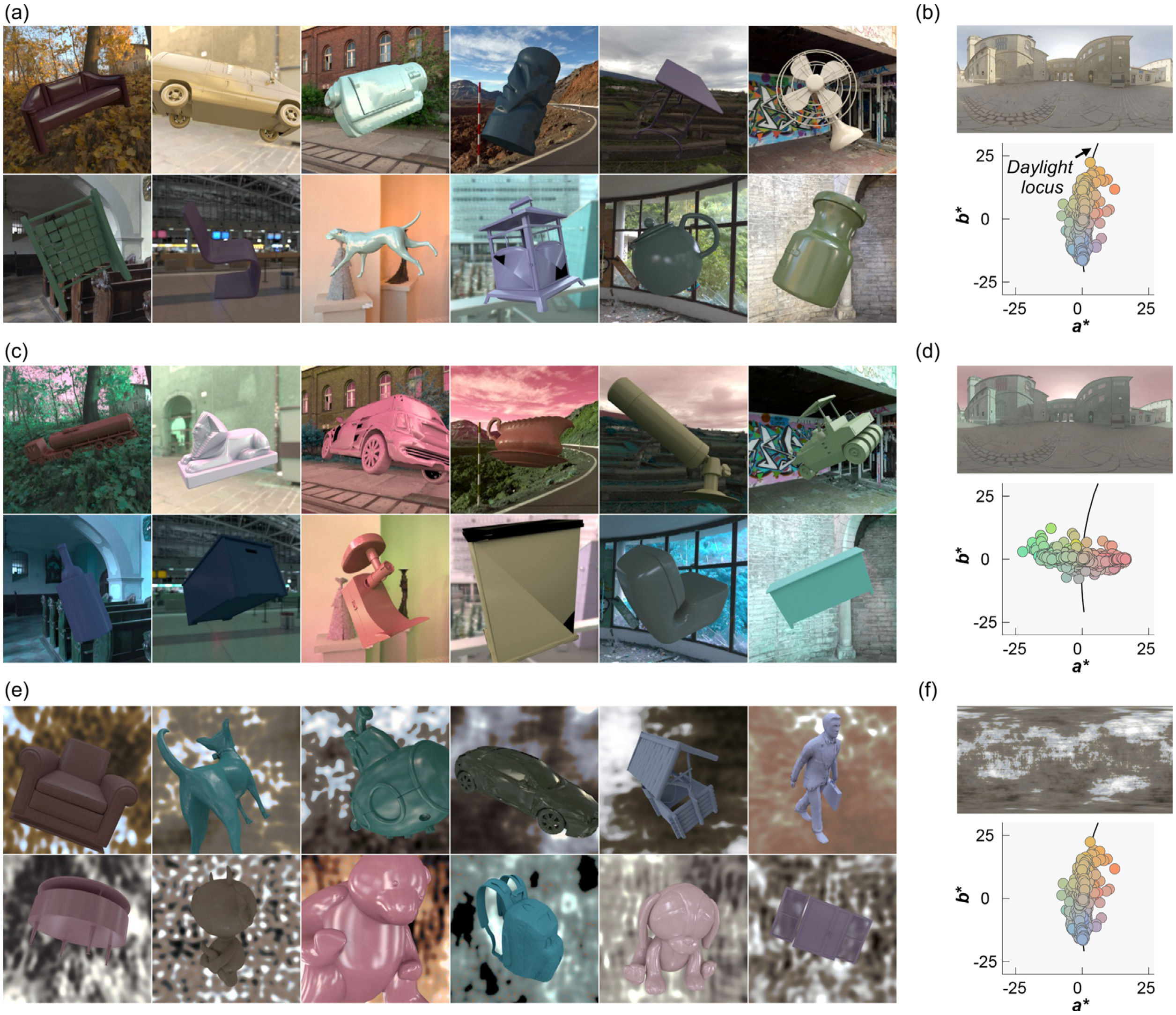

Figure 5. Thirty-six test images (a, c, and e) and an example lighting environment along with its a^*b^* chromatic distribution (b, d, and f) in Experiment 1. (a, b) natural lighting environment, (c, d) +90° gamut-rotated lighting environment and (e, f) phase-scrambled lighting environment. Each test image contains a single object whose shape (and orientation), color, and gloss level were randomly assigned.

explore the full range of each parameter to familiarize themselves with the range of each parameter. After the practice trials, participants adapted for one minute to 20-Hz random-dot dynamic color noise whose mean chromaticity was equal to the chromaticity of equal energy white, then the experiment began. Initial physical reflectance parameters for the reference object were randomized for each trial. The viewing distance was kept constant at 49 cm from the LCD monitor. Participants signaled when the matching was completed by a button press and the selected underlying physical reflectance parameters of the reference object were recorded as the participant's response.

In Experiment 1, there were 36 test images and 2 control images that contained the same bumpy sphere as the reference objects presented under the reference lighting environment (symmetric matching). The reflectance parameter in control images was fixed at hue 120, chroma 22, lightness 40, and Pellacini's c 0.1305 for the control image 1, and hue 320, chroma 14, lightness 60, Pellacini's c 0.0152 for the control image 2. One session thus consisted of 38 settings, and all participants completed three sessions in total. One setting took 66.4 ± 15.5, 53.5 ± 13.0, and 41.1 ± 7.9 seconds (mean ± 1.0 standard deviation across 10 participants) for sessions 1, 2 and 3, respectively. In Experiment 2, one session consisted of 216 trials (216 test images), and 2 sessions were completed. One setting lasted 11.5 ± 3.23 and 8.42 ± 2.18 seconds (mean ± 1.0 standard deviation across 10 participants) for sessions 1 and 2, respectively. For both experiments, there was a break between sessions.

Experiment 1

Results

Control condition

We first show participants' settings for the control image 1 (symmetric matching) in Figure 6. It is clear that participants can highly accurately set the reflectance parameters close to ground-truth values when the shape and lighting environment are identical between test and reference images. The accuracy of the setting was similar for the control image 2. This finding is not surprising, but these error values can be usefully taken as a measure of matching precision for each parameter.

Main conditions

Figure 7 shows the main results for Experiment 1. Each data point shows the averaged setting across 10 participants for 1 test image. From left to right, each subplot shows results for hue, lightness, chroma, and Pellacini's c, respectively. Black numbers at the upper left corner of each subpanel show the correlation coefficient between ground-truth value and participants' settings, calculated separately for each observer first and averaged across 10 participants.

Note that, in an asymmetric matching experiment, participants' settings should be interpreted relative to the reference image used in this study. Taking the Pellacini's c as an example, data points falling exactly on the diagonal unity line shows that participants set the Pellacini's c of the reference object so that it is the same as the test object. This means that the perceived gloss level of the test object and reference objects are the same when both objects have the same Pellacini's c. In an alternative case, data points located below the diagonal line mean that those test objects would appear to be less glossy than the reference object if the test objects' Pellacini's c (ground-truth) were applied to the reference object (which is why participants needed to decrease the Pellacini's c of the reference object to achieve matching). For this reason, absolute error values are likely to vary by the choice of reference image, and thus we evaluated the accuracy of the settings using correlation coefficients between human settings and ground-truth values, which are not affected by additive shift or multiplicative scaling of the data points, which might have been introduced by the choice of the reference image. This relative measure also allows comparisons across different parameters.

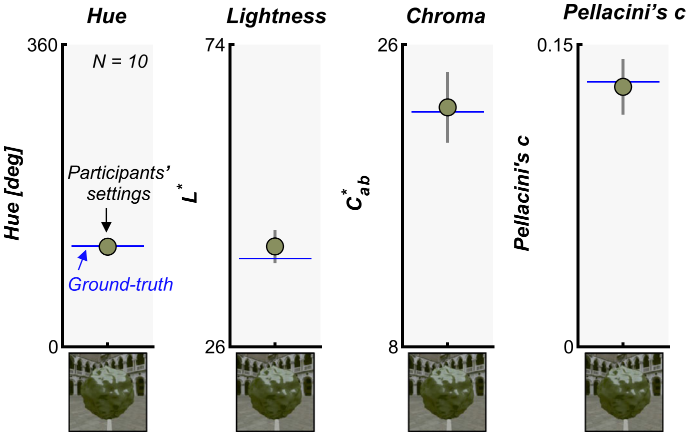

Figure 6. Results for the control condition (control image 1), where the test image and reference image had identical shapes and lighting environments to measure the precision of participants' settings along each parameter. The settings were first averaged across 3 sessions for each observer and then averaged across 10 participants. The image on the *x*-axis shows control image 1. The blue horizontal line depicts the ground-truth values assigned for the test object. The error bars show ±1.0 standard deviation across 10 participants.

At first glance, the data points in Figure 7 show large scatter, except for hue settings, and the data seem noisy. However, we emphasize that these settings were highly consistent across participants as illustrated in Figure 8, meaning that all participants systematically showed similar deviations from ground-truth. The left bars show the averaged correlation coefficient between human settings and ground-truth (the same as the black numbers in the upper left corners of plots in Figure 7), and the small circles show individual participants. The right bars show the averaged correlation among participants (interparticipant correlation). To compute this, for each participant, we calculated the correlation coefficient between the participant's settings and settings averaged across all participants (Nili et al., 2014). This reveals that human-human correlations are generally high regardless of human–groundtruth correlation, making it clear that the scattered data patterns in Figure 7 are not due to noise.

Returning to Figure 7, hue settings in natural environments (Figure 7a) show a significant correlation with ground-truth values, indicating that hue judgment is stable regardless of test lighting environments. This observation holds well for phase-scrambled environments (Figure 7c) and less so for gamut-rotated environments (Figure 7b). This generally high degree of hue constancy may not be very surprising because the circular correlation between mean hue value over the object region and ground-truth values across 36 test images was 0.92. This means that basing a judgment on the mean hue value over the object region (Milojevic, Ennis, Toscani, & Gegenfurtner, 2018) in each trial would lead to high correlation with ground-truth, but this might also suggest that hue is physically a relatively stable quantity, at least for the environmental illuminations we tested here. To evaluate

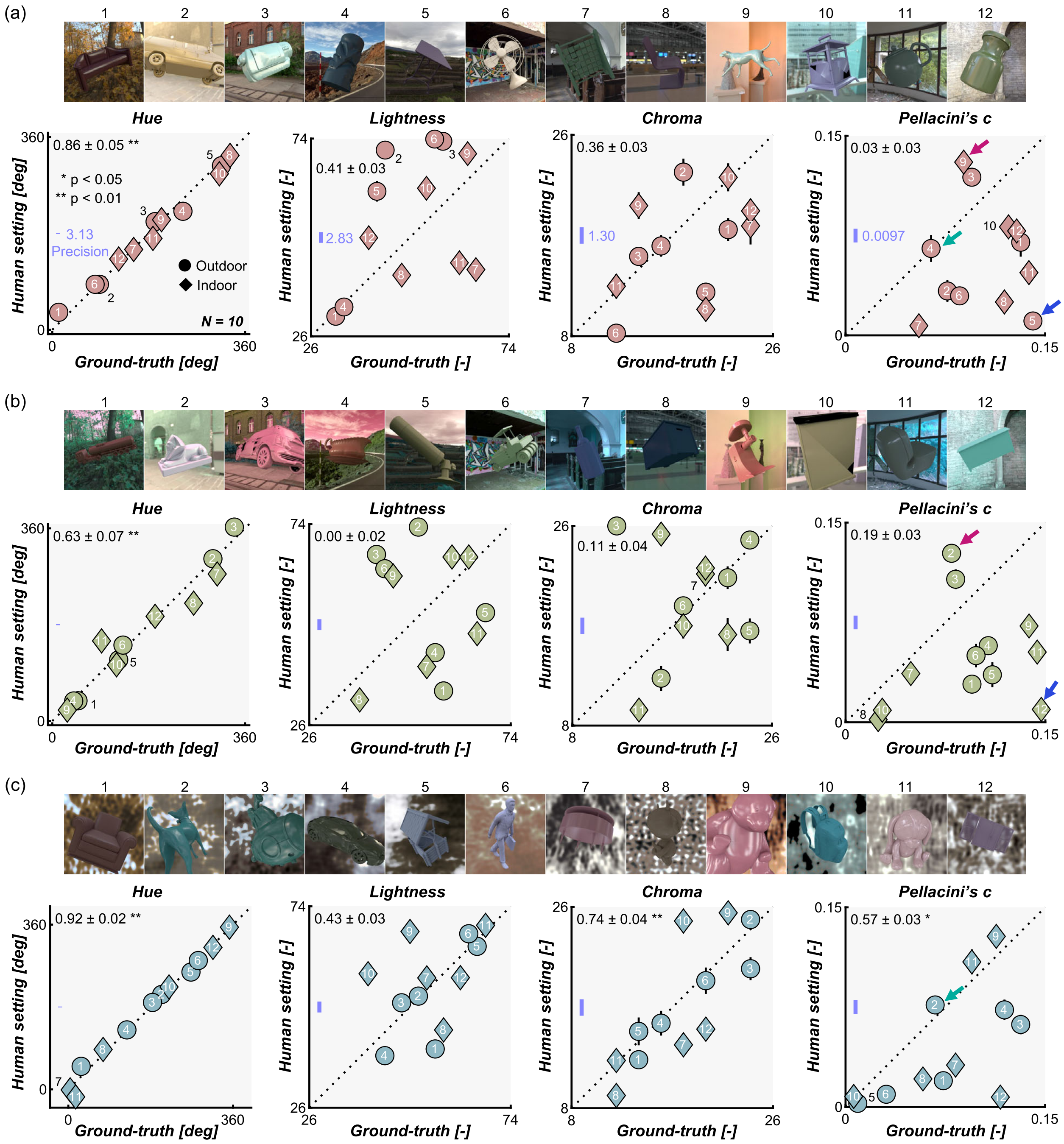

Figure 7. Results in Experiment 1 for (a) natural lighting environments, (b) gamut-rotated lighting environments, and (c) phase-scrambled lighting environments. Each setting was averaged across 10 participants. Error bars show ±1.0 standard error (SE) of the mean over 10 participants, which is smaller than the data point for most cases. Upper images in each panel show test images, with numbers showing the correspondence to data points. The number at the left upper corner in each subpanel shows the correlation coefficient (mean ± 1.0 SE) between ground-truth values and participants' settings calculated separately for each participant and averaged across 10 participants. The blue line represents the precision computed from the mean absolute error in the symmetric matching data. Small colored arrows show test images whose gloss level was judged to be particularly high (red arrows), low (blue arrows), and the same level as the reference image (green arrows).

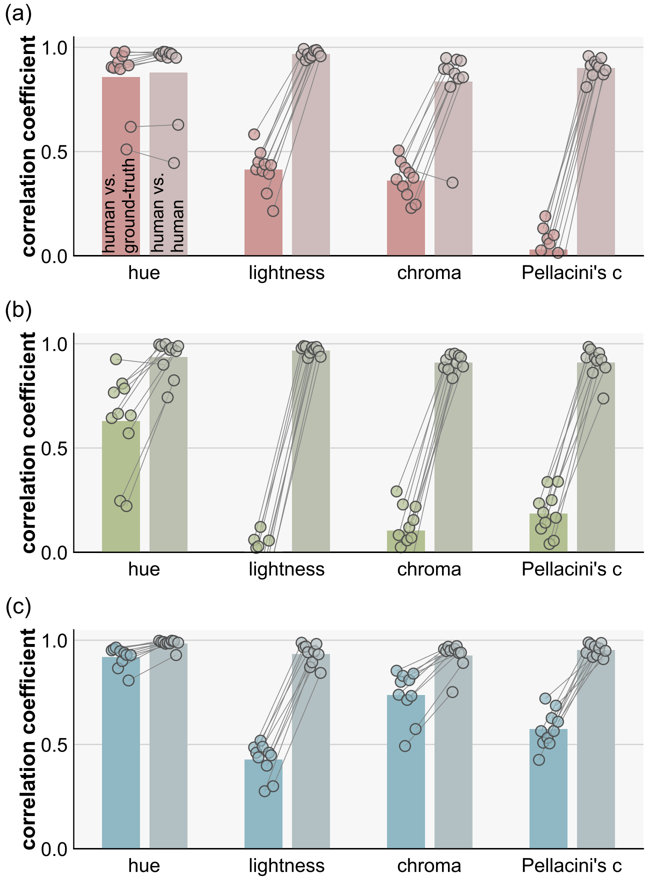

Figure 8. Left bars and data points show correlation coefficients between each human participant and the ground-truth and right bars and data points show interparticipant correlation (correlation between each participant and the average across all participants) for (a) natural, (b) gamut-rotated, and (c) phase-scrambled lighting environments. Each dot shows an individual participant and each bar shows the average across 10 participants.

the influence of the type of lighting environment (natural, gamut-rotated, and phase-scrambled) on the correlation coefficient between human settings and ground-truth, we performed one-way repeated measures analysis of variance, which confirmed a significant effect of illuminant type, $F(2,18) = 11.9$, $p = 5.14 \times 10^{-4}$. Post hoc multiple comparisons (Bonferroni's corrected $p = 0.05$) showed significantly higher correlation for natural environments (Figure 7a) than gamut-rotated environments (Figure 7b) and phase-scrambled (Figure 7c) than gamut-rotated environments (Figure 7b), suggesting worse hue constancy in chromatically atypical lighting environments ,which implies the role of a daylight prior in judging the illuminant influence (Pearce et al. 2014; Weiss, Witzel & Gegenfurtner, 2017).

In contrast, lightness and chroma settings are scattered in a disorderly way, leading to generally lower correlation with ground-truth than for hue settings. The correlation coefficient was statistically significant only for the chroma setting in (c) phase-scrambled lighting environments. Similarly, settings of Pellacini's c were not highly correlated with ground-truth though the correlation is significant for (c) phase-scrambled environments. Here, most data points fall below the diagonal unit line, meaning that the perceived gloss level of test objects was generally lower than the reference object. This might reflect either the more even sampling of surface normals for the reference geometry, or a tendency of the Uffizi probe to make objects appear to be particularly glossy.

To check the presence of a learning effect over three sessions, we computed the correlation coefficient between human settings and ground-truth values. We found that for hue setting the correlation coefficient over 36 images significantly increased from session 1 to session 2 (Bonferroni's corrected $p < 0.05$). However, no learning effect was observed for chroma, lightness and Pellacini's c, which is unsurprising as there was no feedback on how close each setting was to ground-truth value.

Discussion

Subsequent subsections discuss potential underlying reasons for highly consistent error patterns for chroma, lightness, and gloss judgments.

Interaction between diffuse reflectance and specular reflectance

First, we considered the possibility that specular reflections may have contaminated the perception of body color. If so, we should observe that deviations between lightness and chroma settings by participants and ground-truth values become larger as a function of physical (or perceived) gloss level. Accordingly, we computed correlation coefficients between the ground-truth Pellacini's c and lightness error (human setting minus ground-truth) and between the ground-truth Pellacini's c and chroma error across 12 test images, separately for each type of lighting environment. However, we found no significant correlation under any type of lighting environment. We repeated this analysis using Pellacini's c set by participants instead of the ground-truth Pellacini's c, but again there was no significant correlation under any type of lighting environment. Thus, large errors in chroma and lightness observed in Figure 7 are unlikely to be due to masking or intrusion by the specular reflection.

Similarly, we next asked whether diffuse reflection contaminated the perception of gloss. In other words, were participants more likely to make errors in gloss settings for a certain range of body colors? Figure 9 visualizes the magnitude of errors (Pellacini's c set by participants minus ground-truth Pellacini's c) as

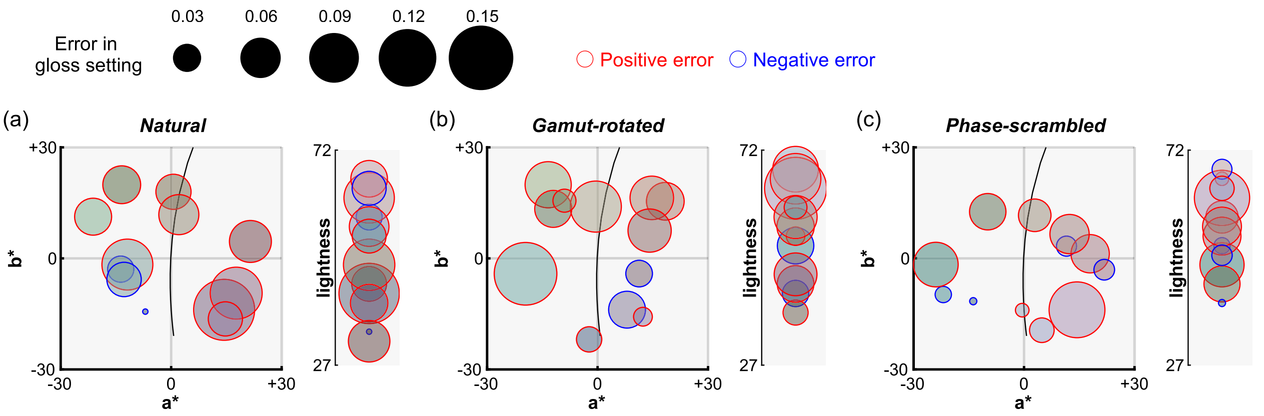

Figure 9. Analysis of interaction between body color (diffuse reflectance) and perceived gloss level for (a) natural, (b) gamut-rotated, and (c) phase-scrambled lighting environments in Experiment 1. Each data point corresponds to one test image and its color shows the body color (in sRGB format). The size of each data point represents the error of gloss setting (human setting minus ground-truth). Red and blue edge colors show positive and negative errors, respectively. Left plot shows representation in *a*b** chromatic plane while the right plot shows lightness axis. There was no systematic trend found between the body color and the magnitude of gloss errors.

a function of body color. However, we found no significant correlation between errors in gloss settings and each of the color parameters (i.e., hue, chroma and lightness) under any type of lighting environment after the correction of significance level (by Bonferroni's correction), showing that there are no noteworthy interactions between failures of gloss constancy and body color.

Overall, these analyses confirmed that error patterns in color settings were independent from specular reflectance. Similarly, errors in gloss judgements were not systematically affected by diffuse reflectance. Evidently judgments of the components of reflectance are based on distinct image information or internal representations, not a single composite representation of the entire bidirectional reflectance distribution function (Nicodemus, Richmond, Hsia, Ginsburg, & Limperis, 1997).

Interaction between lightness and chroma

We wondered whether there might be an interaction between parameters related to color, which affected participants' perceptual judgment and matching. One such candidate was the interaction between chroma and lightness as a past observation suggested that perceived saturation is relatively well predicted by C^*_{ab}/L^* (Fairchild, 2013; Schiller, Valsecchi, & Gegenfurtner, 2018). In other words, it is possible that participants judged the chroma and lightness match by simply judging the match in perceived saturation between test and reference objects. If that's the case, ground-truth values and participants' settings when represented in C^*_{ab}/L^* should correlate well. As shown in Figure 10, we found significant correlations for all three types of lighting environment. This explains at least partially the scattered setting patterns for chroma and lightness in Figure 7, especially for the phase-scrambled environments (Figure 7c), although a substantial portion of the variance remains unaccounted for with the natural (Figure 7a) and gamut-rotated environments (Figure 7b).

Image statistics

Another candidate account would be that participants based their judgements on various summary statistics directly accessible from test images. Indeed, since there is no direct way for participants to access the ground-truth values of gloss and body color, such a strategy could be a reasonable alternative approach to performing the task. To predict participants' lightness and chroma settings, we calculated mean, median, standard deviation, skewness, kurtosis, first quartile (Q1), third quantile (Q3), and minimum and maximum values of lightness and chroma across the object region in each test image. Surrounding context was excluded from the computation.

Additionally, to understand why gloss constancy was poor in our task, we selectively looked at test images that were rated particularly high gloss (discrepancy from ground-truth, +0.0411, +0.0469) and low gloss (discrepancy, −0.129, −0.136), as well as images where ground-truth and human settings well matched. Figure 11a shows example images (these images are labeled by colored arrows in Figure 7). It is evident that specular reflection patterns are visibly different across images. Objects in high-gloss images (surrounded by red squares) seem to receive strong directional lights in the environment and consequently have a readily visible specular reflection pattern though physically the specularity is around the middle of the range (Figure 7). In contrast, objects in low gloss images (surrounded by blue squares) are placed under a

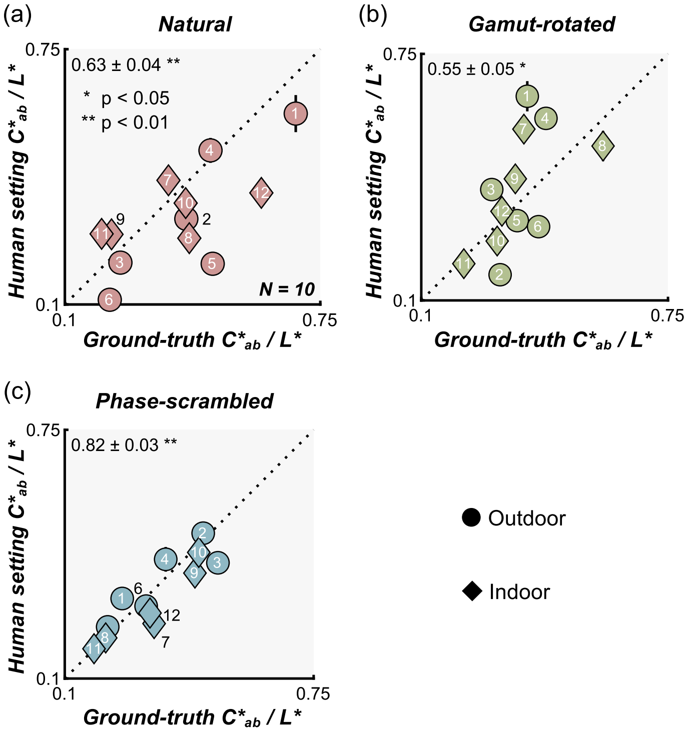

Figure 10. Scatter plot to see whether interaction between chroma and lightness (saturation defined as C^*_{ab}/L^*) can explain the observed chroma and lightness constancy failures. The x-axis shows ground-truth saturation while the y-axis shows saturation computed from matching results. Significant correlations between ground-truth and human settings suggest that participants might have used the perceived saturation to help determine whether the color was matched between reference object and test object.

dim and cloudy lighting environment or indoor scene. Specular reflection is present in these cases too, but it spreads across the surface and that is why, when mixed with diffuse images, the specular reflection patterns are hard to detect even though both objects have nearly the highest possible specular reflectance. Finally, for the images where human settings and ground-truth were well-correlated, specular reflections are moderately visible. In sum, it visually makes sense that these images were rated by participants in these ways, which encouraged us to compute image metrics based on patterns of specular reflection.

Thus, in addition to basic descriptive statistics, we used metrics computed from the structure of the specular reflection image (Marlow, Kim, & Anderson, 2012; Schmid, Barla, & Doerschner, 2021) to predict participants' gloss settings. As shown in Figure 11b, we first converted the original image to a luminance image in which each pixel has the unit of candela per square meter, masked out the object region, and subtracted the diffuse component from the image, which resulted in test images with specular reflection alone. Then, we extracted pixels whose intensity is higher than k% value of the highest intensity across this specular image, where k took the following values 0, 1, 3, 5, 10, 20, and 40 to get rid of the region of specular reflection that stems from secondary and higher-order inter-reflections. Using this thresholded highlight image we calculated the following three metrics. The first metric is coverage, corresponding with the proportion of area covered by the highlight relative to the whole object area as depicted in the top left part of Figure 11c. Second, we calculated the sharpness. Using a spatial convolution, this metric emphasizes the region where luminance rapidly changes and sharpness is defined as a mean value of the convoluted sharpness map (Vu, Phan, & Chandler, 2012) as shown in top right part of Figure 11c. For coverage and sharpness, model predictions were affected by the cut-off percentage to threshold the highlight regions and thus we selected an optimal value of k that produced the highest correlation value with human settings. We note that, by searching for the optimal cut-off threshold in this way, we considered a possibility that a "low-light" region of the specular image could contribute to human gloss percept (Kim, Marlow, & Anderson, 2012). Finally, the third metric was contrast, which essentially measures the spatial luminance variation over the surface. The standard way would be to calculate a contrast from the raw highlight image directly. However, considering a previous observation that perceived gloss is affected by the modulation of a specific frequency channel (Boyadzhiev, Bala, Paris, & Adelson, 2015), we first decomposed the raw highlight image into eight sub-band images using a Gaussian band-pass filter (upper and lower cut-off frequencies: 1.5 to 3.0, 3.0 to 6.0, 6.0 to 12.0, 12.0 to 24.0, 24.0 to 48.0, 48.0 to 96.0, 96.0 to 192, and 192 to 384 cycles/image) and a subset of sub-band images are shown in the lower part of Figure 11c. Here, the highest center frequency was 18.2 cpd, which would be comfortably resolved for participants with the visual acuity 20/20. We calculated the RMS contrast, equivalent to the standard deviation of the pixel intensities, for each sub-band image as well as for an aggregated image across all frequencies. This means that, unlike coverage and sharpness, which has one parameter, the sub-band contrast metric has two free parameters (i.e., the cut-off pixel intensity and the cut-off spatial frequency band), and optimal values producing the highest correlation with human settings were selected. For all three metrics, searching for the best parameters was performed separately for each type of lighting environment (natural, gamut-rotated, phase-scrambled).

Each bar in Figure 12 shows the absolute values of correlation coefficient between image statistics and human settings for a natural lighting environment (Figure 12a), gamut-rotated lighting environment (Figure 12b), and phase-scrambled lighting environment (Figure 12c). Higher values indicate that the models

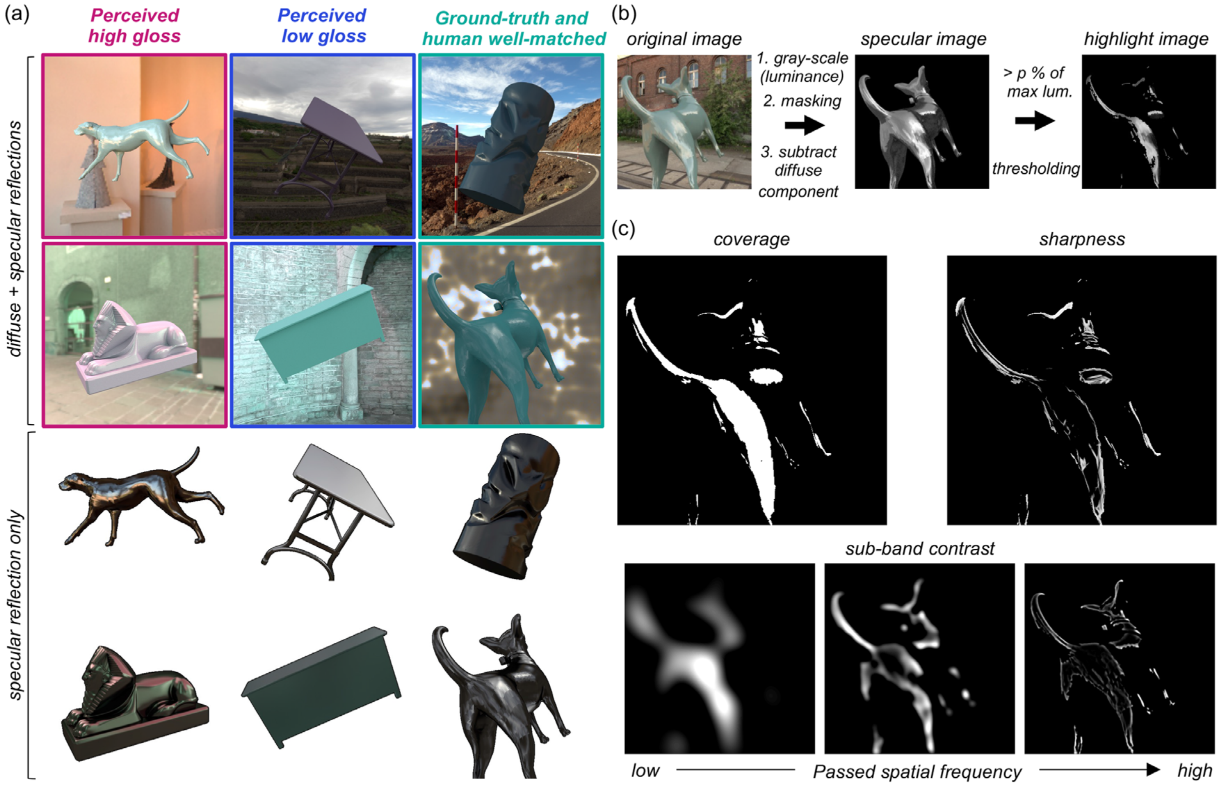

Figure 11. (a) Example images where humans perceived high gloss (left column), low gloss (center column), and where ground-truth gloss level and human settings matched well (right column). The top six images show original images where diffuse and specular reflections are both included. The bottom six images show objects without diffuse reflection and surrounding context to visualize the spatial structure of specular reflection on which participants might have based their gloss judgements. (b) Process to convert original test image to thresholded highlight image. (c) Calculation of three metrics to predict human gloss percepts: coverage, sharpness and sub-band contrast (see the main text for details).

capture participants' settings better. The magenta shaded region shows the interparticipant noise ceiling, between upper bounds (same as the magenta values in Figure 7) and lower bounds, computed by a similar procedure as the upper bound, but by calculating the correlation coefficient between the left-out participant's settings and settings averaged across all other participants. This range effectively defines a bound on how well any image-computable model could perform. These image statistics were computed from test objects that include both diffuse and specular reflections. For lightness and chroma, we also considered image statistics computed directly from the diffuse component only (without specular reflection) to test the idea that humans might have effectively discounted the specular reflection from the test image. The red diamonds show the correlation between image statistics calculated directly from the diffuse image and participants' settings. If participants took such a strategy, the red diamonds should come higher than the green bars.

For lightness and chroma, it is clear that, although the correlation between human settings and ground truth are remarkably low (as shown in Figure 7), participant settings are highly correlated with simple statistics, such as mean lightness and mean chroma over the object region, nearly touching the noise ceiling level in some cases. Interestingly, for lightness, the maximum luminance of diffuse components (red diamonds) predicts human settings better than maximum luminance of the original image (bars), consistent with a past observation (Giesel & Gegenfurtner, 2010).

Looking at Pellacini's c, overall image statistics correlate with human settings less than for chroma and lightness, and rather surprisingly the maximum luminance values are generally good predictors for any type of lighting environment. Yet, there are still significant disparities between the predictors and human performance. Thus, these models explain the failure of gloss constancy to some extent, but not enough to capture the complexity of gloss perception. Furthermore, we also checked whether any of the light map statistics shown in the luminance histogram diagram in Figure 4 could predict the gloss percept. However, we found no significant correlation between any of the statistics and human gloss settings. Importantly, these light map statistics are computed for the whole light map, whereas the object in each test image receives lights from only

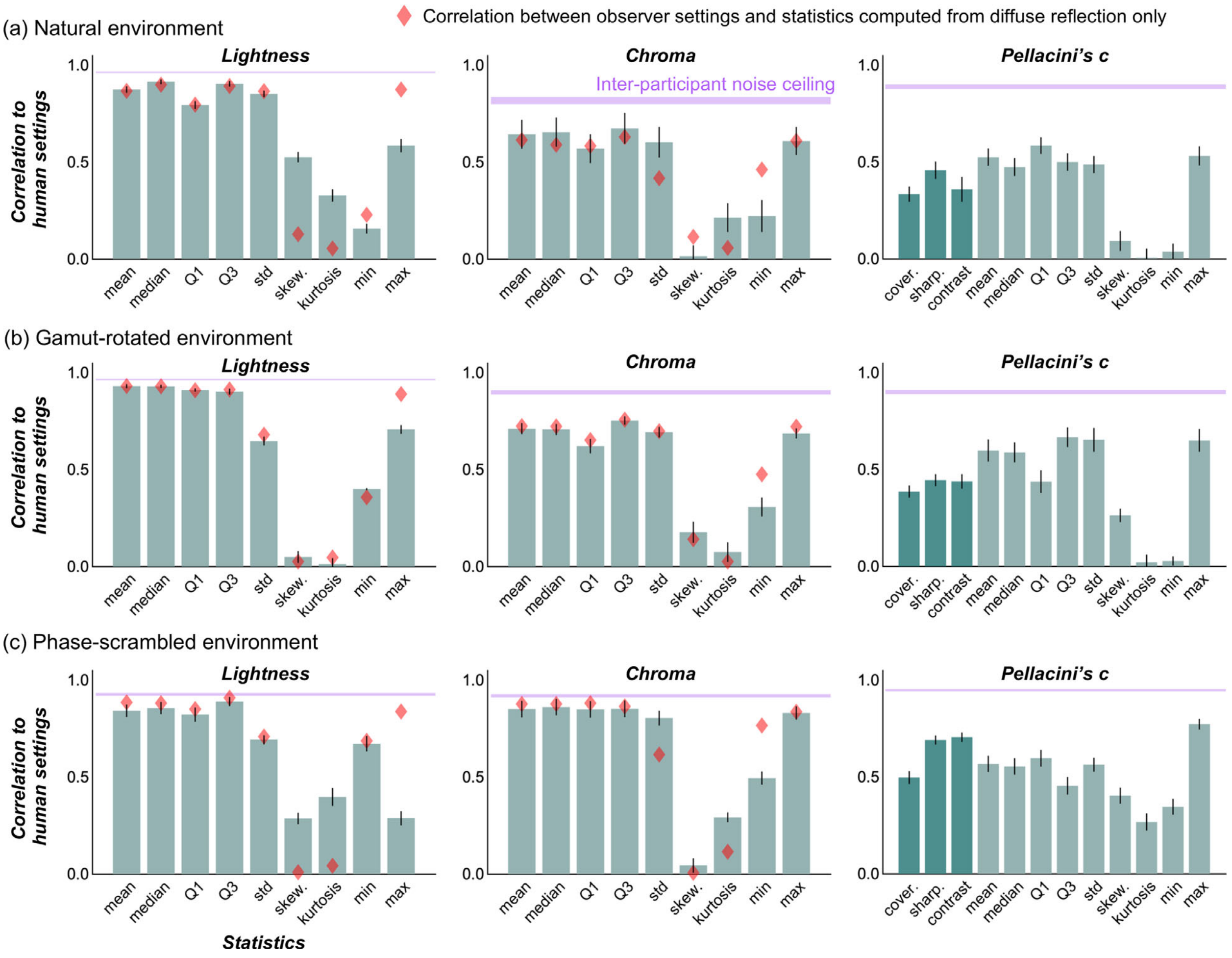

Figure 12. Correlation coefficients between image statistics (calculated over the object regions in test images) and human settings. (a) Natural environment, (b) gamut-rotated environment, and (c) phase-scrambled environment. The magenta shaded areas show an interparticipant noise ceiling, equivalent to the correlation across participants, which correspond to magenta numbers in Figure 8. Red diamonds show the correlation between image statistics calculated over the object region in test images that had only diffuse reflection (no specular reflection).

a limited part of the light map. Thus, if instead we widely changed the viewpoint and measured perceived gloss for each viewpoint, we might have seen a higher correlation.

To summarize the results from Experiment 1, both success and failures of color constancy were well-captured. Hue constancy holds remarkably well. Failures of chroma and lightness constancy were systematic and largely predicted by simple metrics such as mean chroma or mean lightness over the object surface. Gloss constancy was very poor for our stimuli, but matches were highly consistent across participants, and the simple image statistics we investigated explained human behavior to a limited extent.

To grasp the complex nature of gloss perception better, we felt that the 36 test images used in Experiment 1 may not be enough. Thus we conducted a follow-up Experiment 2 with 216 test images using a factorial combination between 12 lighting environments and 18 shapes, and the perceived gloss was again measured using the asymmetric matching task.

Experiment 2

Test images and procedure

We generated 216 images using 18 shapes sampled from those used in Experiment 1, and 12 lighting environments (Figure 13). The hue, lightness, and chroma of test objects and matching objects were fixed at 188.2°, 33.9, and 11.9, respectively. We picked this greenish color because we speculated that the visibility of the specular reflection would be higher when the body color falls on the axis orthogonal to daylight locus (i.e., red–green axis); as seen in Figure 4, the colors of the specular reflection are distributed mainly along

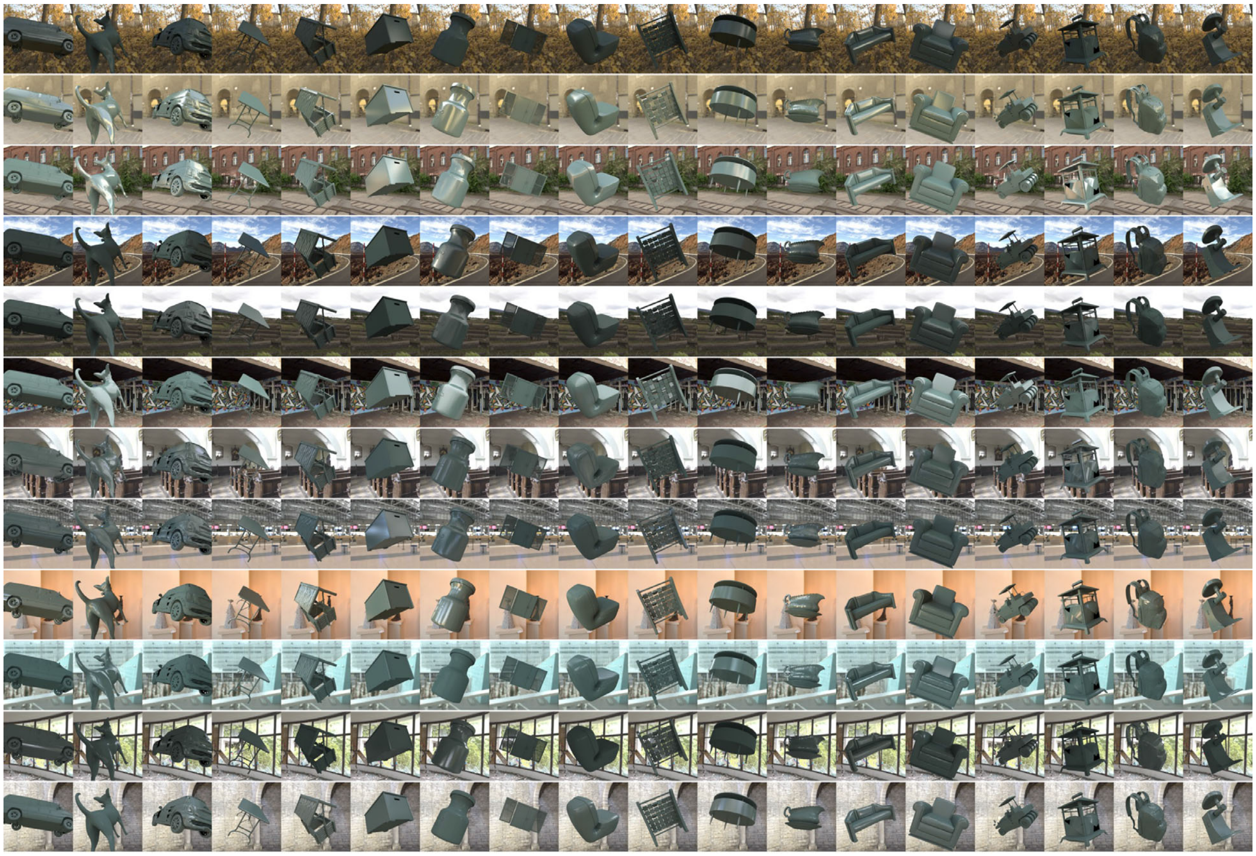

Figure 13. The 216 test images used in Experiment 2. Images were generated by the factorial combination between 12 natural lighting environments in Experiment 1 and 18 shapes that were randomly sampled from the 36 shapes in Experiment 1. Body color was fixed as a greenish color, which we found useful to increase the visibility of specular reflection on the objects' surface.

the blue–yellow axis. The task of participants and procedure were identical to Experiment 1, except that participants' adjusted only Pellacini's c in Experiment 2 and the color of the reference object was fixed to the same green as the test object. For each test image, we applied a random Pellacini's c ranging from 0 to 0.224, which corresponds with the range from 0 to 0.0999 in Ward's specularity. The maximum value of Pellacini's c here differs from the value in Experiment 1 (0.149) because the lightness value of the object for the conversion between Pellacini's c and Ward's specularity was 33.9 instead of 50. One session consisted of 216 trials, and all participants completed two sessions in total.

Results and discussion

Figure 14a shows participants' settings (averaged across 10 participants) in Experiment 2 grouped by the test lighting environment. Each data point corresponds to one shape, and thus there are 18 data points in each subplot. Globally, looking through subplots, it is evident that data points deviate substantially from the unity line (diagonal dotted line), showing that human settings and ground-truth strongly disagree. It is also noticeable that the slope of the red regression line differs from one lighting environment to another (minimum of 0.34 and maximum of 0.90). A higher slope means that on average under that lighting environment objects appear to be more glossy. Also, the correlation coefficients between human settings and ground-truth values (black numbers at the top left) were significant in all lighting environments, unlike in Experiment 1, but the values vary from 0.46 to 0.78 showing that the variability owing to the objects' shape also differs from one lighting environment to another. Thus, simple transformations are unlikely to equate the perceived gloss level in one lighting environment to another environment. This finding is inconsistent with past research using smooth spheres (Doerschner et al. 2010; Fleming, Dror, & Adelson 2003), and the use of a variety of shapes in this study is likely to be a reason. Overall, these results suggest that perceived gloss somewhat correlates with the underlying physical specular reflectance, but also differs owing to the change

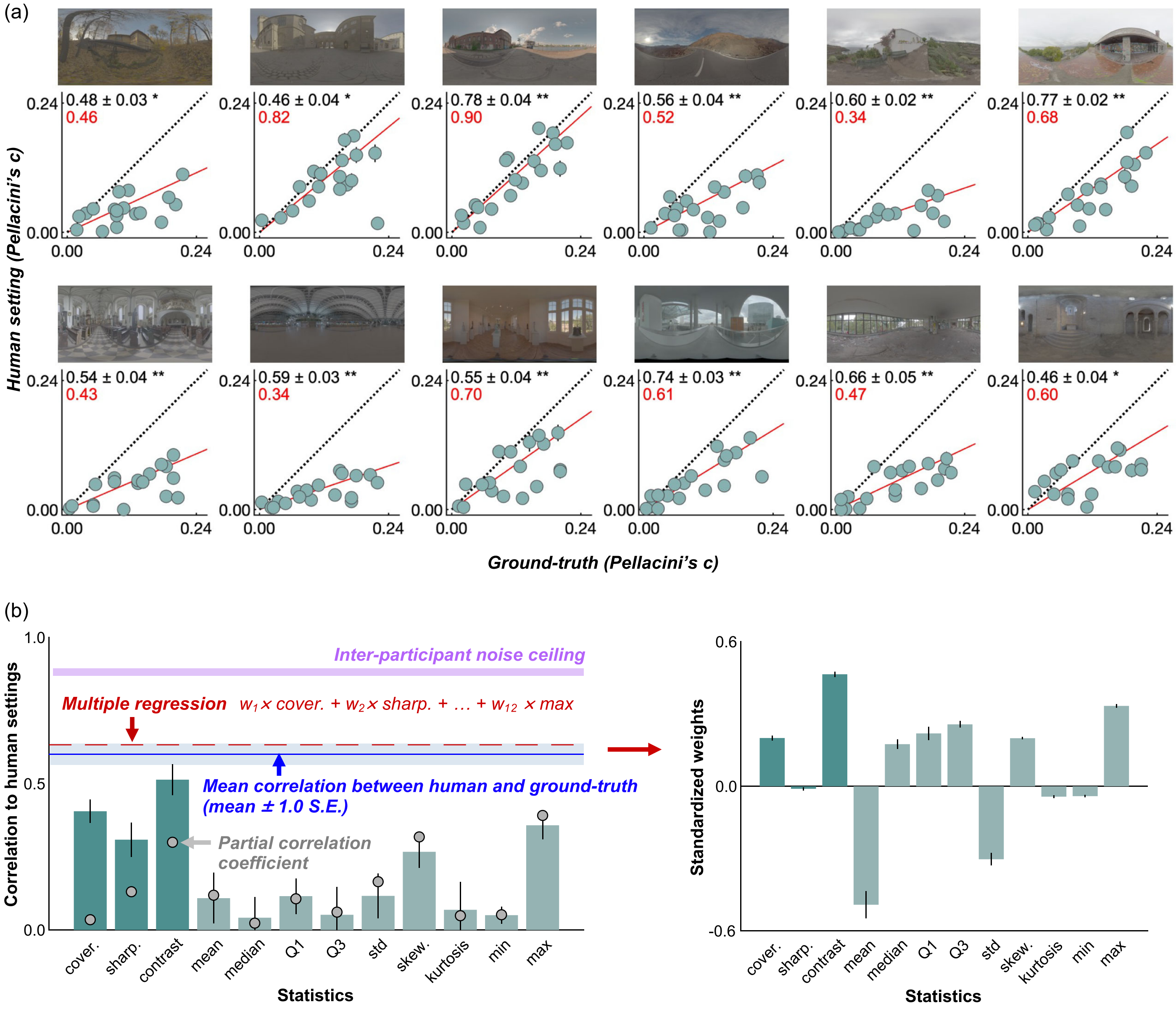

Figure 14. (a) Average setting across 10 participants in Experiment 2, where settings were grouped by lighting environment. Each data point denotes one shape, and thus there are 18 data points in each scatter plot. Error bars show ±1.0 standard error (SE) across 10 participants. Regression lines fitted to the 18 data points are shown by a solid red line. Black numbers in the upper left corners of the plots show the correlation coefficient (±1.0 SE) between human settings and ground-truth, computed for each participant first and then averaged across 10 participants. Red numbers show the slope of the regression line. Higher slope means that, on average, that lighting environment produced a high degree of perceived gloss. (b) Lefthand bar plot shows correlation coefficient between image statistics and human settings. The dark green bars show image statistics computed from specular reflection images, and the light green bars show simple luminance image statistics. Error bars show ±1.0 standard deviation across 10 participants. The blue line shows the correlation coefficient between participants' settings and ground-truth values, averaged over 12 lighting environments (average over 12 black numbers in panel (a)). The pink shaded areas show an interparticipant noise ceiling computed in the same way as for Experiment 1 (Figure 12), which no model can exceed. The gray filled circles show the partial correlations between each image statistic and human settings, removing the influence of the correlation between image statistics and ground-truth. The red line shows the correlation coefficient between human settings and a weighted sum of 12 statistics where weights were optimized through multiple regression. The right-hand bar plot shows a standardized regression coefficient (Borgonovo & Plischke, 2016) for each statistic averaged across 10 cross-validations, and error bars show ±1.0 standard deviation.

of lighting environment, showing failures of gloss constancy.

We next asked whether image statistics explain these human behaviors. The lefthand bar graph in Figure 14b shows the absolute values of correlation coefficient (averaged over 10 participants) between each statistic and human settings over 216 images. Unlike the results of Experiment 1 (Figure 12), basic image statistics showed a fairly low correlation. The maximum luminance shows lower correlation than Experiment 1, and thus the high effectiveness of the maximum luminance in Experiment 1 was presumably because of the small number and limited variation of images used in Experiment 1. In contrast, the sub-band contrast metric showed a relatively high correlation. However, this correlation was significantly lower than the correlation between human settings and ground-truth shown by the blue line (two-tailed paired t test), $t(18) = 3.25$, $p = 0.0045$. Moreover, it is worth noting that this contrast metric (and others) also correlates with the ground-truth value to some extent, and this factor may be a part of the reason why these metrics correlated with human settings. Thus, to remove the influence of ground-truth, we additionally computed a partial correlation coefficient between the image statistics and human settings (gray filled circles), removing the correlation between image statistics and ground-truth values. We found that they are lower than the original correlation coefficient, especially for image metrics computed from specular reflections (left three gray points). This result suggests that these models correlate well with human settings mainly simply because they correlate with ground-truth, not necessarily capturing systematic failures of gloss constancy consistently exhibited by participants. For other simple image statistics, partial correlation and original correlation coefficients are similar to each other, likely because those simple metrics do not correlate with the ground-truth values well. Finally, we tested what happens if we assume observers combined multiple statistics instead of using a single cue independently. To test this idea, we ran multiple regression analyses using all 12 metrics as independent variables to explain the participants' settings. When we did this, we trained a regressor using the averaged settings across 9 participants to find optimized weights for 12 metrics, and then using these optimized weights we calculated the correlation between the regressor's prediction and the settings by the left-out participant. We repeated this 10 times (leave-one-out cross-validation), and the average across the 10 correlation coefficients is shown by the horizontal dashed red line. The right-hand bar plot shows standardized weights for each statistic, averaged across 10 validations. However, the improvement owing to the integration of multiple cues was marginal, and consequently there is still much room between this level and the noise ceiling level. We also looked at the correlation between human settings and each of the statistics of lighting environments shown in Figure 4, but the correlation coefficients were overall low, and the highest value was 0.11 for skewness. Thus, these statistics are not predictive of human gloss perception in Experiment 2, being consistent with the observation in Experiment 1.

General discussion

How do we overcome huge variations in the proximal image to create a stable percept of the color and gloss of objects? To address this question, we measured color and gloss constancy together using an asymmetric matching task under a diverse set of lighting environments. Our results revealed a strong asymmetry across hue, chroma, and lightness constancy; the degree of hue constancy was generally high, although it slightly decreased when the lighting environment had atypical chromatic properties, whereas lightness and chroma constancy were in general severely limited (except for chroma settings under phase-scrambled environments). These failures of chroma and lightness constancy were well-captured by the saturation metric (chroma/lightness) and simple image statistics over the object region in the image. In contrast, gloss constancy was generally poor (i.e., gloss ratings depended on lighting environment and shape), but phase-scrambling directional lighting geometry did not impair gloss constancy additionally. Image statistics explained those failures of gloss constancy only to a limited extent. One major finding in this study is that, although there have been observations that simplistic image metrics can account for large variations in human gloss perception, when the diversity of shape and lighting approaches that seen in real-world environments a significant amount of consistent variance in gloss judgments remains unexplained.

This study presented constancy errors that were remarkably consistent across tested participants, but there have been reports that the degree of color constancy could vary substantially from one experimental condition and paradigm to another (Foster, 2011). Our findings on the limited gloss constancy are in fact largely in line with findings from several past studies. Pont and te Pas (2006) reported that participants showed significant failures of constancy in a task where two presented spheres illuminated differently have either the same reflectance properties or not. Nishida and Shinya (1998) used a reflectance matching paradigm using both albedo and specular reflectance and showed that participants' lightness and glossiness judgment were heavily influenced by object shape. Olkkonen and Brainard (2010) showed systematic failures of gloss constancy due to changes of lighting environment. We note that all of the computational gloss models we implemented in this study ignored the surrounding context. It is possible that a model that incorporates the surrounding context into gloss computation might account better for human behaviors. Also, our stimuli were all static, but adding movement to the stimuli would likely increase the degree of gloss constancy (Dovencioglu, Ben-Shahar, & Barla, Doerschner 2017; Doerschner, Boyaci, & Maloney, 2011; Gigilashvili, Thomas, Pedersen, & Hardeberg, 2021; Scheller Lichtenauer, Schuetz, & Zolliker, 2013; Wendt et al., 2010). In addition, it is worth noting that all these experiments reporting constancy failures—including our own—were conducted using computer monitors, and whether this finding applies to real-world scenarios needs careful investigation. For example, one challenge associated with performing the asymmetric matching paradigm on monitors is limited adaptation to test lighting environments, which is a key contributor in human color constancy (Smithson & Zaidi, 2004). Although the participants were allowed to move their eyes freely during a trial, it is reasonable to assume that the participant looked at the reference image for most of the time during the trial to complete the task. Moreover, images presented on the monitor only occupied a small part of the visual field, and the surrounding region in test images that provides cues to the lighting environments was even smaller (on average, 26.3% of the whole image for Experiment 1 and 18.3% for Experiment 2), which would have made it harder for observers to infer the illuminant influence. However, it is worth noting that the conditions in our experiment were sufficient to enable relatively good hue matches. Finally, we presented only one object in each test image, and presenting multiple objects with various colors in the same scene may have increased the degree of color constancy.

We observed that changing the lighting environment had different effects on different color dimensions. One reason for the superior hue stability could be that the lighting environments we selected did not produce extreme chromatic shifts in the proximal image and consequently pixel hue values did not change severely enough to cause poor constancy. In fact, as shown in Figure 4, although we tried to select environmental illuminations whose color distributions are different from each other, the mean color (shown by a black cross symbol) is still located relatively close to the white point. In contrast, chroma and lightness values shifted enough that mean chroma and lightness values were decorrelated from ground-truth values. However, it is generally true that in natural environments extreme chromatic shifts are rare (Morimoto, Zhang, Fukuda, & Uchikawa, 2022). It is thus interesting to ask how much our selection of lighting environments reflects the true variation of actual lighting environments. If the physical hue values of an object do not vary substantially across scenes in the real world, hue serves as a useful perceptual anchor for object identification under different environments (Ennis, Schiller, Toscani, & Gegenfurtner, 2018; Milojevic et al. 2018).

In this study, we used Pellacini's c to provide the observers with an approximately perceptually uniform scale when adjusting the reference image. This strategy helps to keep precision approximately constant throughout the matching scale. The metric was originally derived from experiments using a grayscale sphere under an area light source and thus naturally shows limited performance outside this context (Gigilashvili & Islam, 2022) as Pellacini, Ferwerda, and Greenberg (2000) themselves acknowledged in their original study. Thus, it could be that the matching scale was not as uniform in our experiment as in Fleming et al. (2003), for example, which—like Pellacini et al (2000)—used spheres. Indeed, we do find a substantial effect of shape on perceived gloss, as has been reported previously (Vangorp, Laurijssen, & Dutré, 2007). It might be argued that the low correlation between human settings and ground-truth for gloss could in some way be influenced by the limited generalizability of the metric to different lighting environments, object shapes, and body color that we manipulated in this study. It is important to note, however, that the design of our experiment does not involve interpreting the settings as absolute perceptual gloss estimates. We are investigating gloss constancy, and thus matches simply provide a measure of the relative apparent gloss between the test and reference. We expect that observers would navigate to very similar points of subjective equality, even if we asked them to adjust a different parameterization of specular reflectance. Changing the parameterization can only shift the points in Figures 7 and 14 along lines parallel to the diagonal in the plots. Yet there is no straightforward remapping that would improve the correlation across all conditions simultaneously radically (only highly implausible nonmonotonic functions of specular reflectance would do this), and thus, at least at a qualitatively level, we do not think that our results depend substantially on the use of Pellacini's c.

An alternative source of error might be that participants could not adjust the surface roughness of the reference object, which is another dimension in Pellacini's model (distinctness, d), which affects perceived gloss. Because the physical roughness was fixed at 0.05 for all reference images and test images in this study, we did not ask participants to adjust the roughness of the reference image. However, we note that it might be possible that surface geometry could have affected the perceived roughness of some surfaces, and for such images the adjustment of roughness dimension might have improved the matching performance.

We showed that a contrast metric computed directly from the specular images showed highest correlation with human settings in Experiment 2, but this model implicitly assumes that humans are capable of separating diffuse and specular reflections from a given image. Such a separation is an ill-posed problem and it is an empirical question how accurately humans can perform the decomposition (e.g., Lee & Smithson, 2017). A complete model of human gloss perception should predict the perceived gloss level from a raw image where diffuse and specular reflections are confounded. A recent effort used deep neural networks trained to output a specular image from a raw image and showed that such a network outperformed a simple alternative highlight detection model based on thresholding and showed relatively high overall similarity to human judgements (Prokott & Fleming, 2022). Another investigation trained unsupervised deep neural networks to model the high-level statistics of images of glossy and matte surfaces, and found that these predicted human gloss judgments better than supervised networks or a range of simpler image statistics (Storrs, Anderson, & Fleming, 2021).

A good perceptual model should reproduce both the successes and error patterns that humans make on an image-by-image basis beyond predicting the overall performance level (Geirhos, Meding, & Wichmann, 2020; Storrs et al., 2021). In this sense, systematic error patterns in Experiment 2 are a potentially useful feature of the dataset as a window into underlying constancy mechanisms. However, we found that our hand-selected features accounted for a limited extent for gloss percepts, and it is a common shortcoming that researchers must select in advance or hand-engineer candidate features. In recent years, big data approaches (often coupled with deep neural networks) have been opening a new avenue to overcome such limitations as networks can learn to extract useful image features by themselves (Liao, Sawayama, & Xiao, 2022; Tamura, Prokott, & Fleming, 2022; Prokott, Tamura, & Fleming, 2021; Sawayama et al., 2022), and this study might also benefit from such approaches. The fact that human judgments can deviate substantially but consistently from ground truth—as we found here—suggests that training a neural network with human responses would potentially yield quite different internal representations than training with ground-truth specular reflectance values. Interpreting the internal representation of the network is a major open challenge, but recent studies have been already tackling the issue (Montavon, Samek, & Müller, 2018). By analyzing the activation of units in the intermediate layers, we might be able to get a hint on specific image features that units in the network highly respond to (e.g., Flachot & Gegenfurtner 2018, 2021). The complexity of the network and interpretability is in a trade-off relationship, and thus using a very shallow network would give a capacity to analyze characteristics of individual units in detail, allowing us to understand extracted features in a meaningful way (Goncalves, & Welchman, 2017; Rideaux & Welchman, 2021).

Keywords: color, gloss, perceptual constancy, directional lighting environments, asymmetric matching

Acknowledgments

The authors thank Alexandra C. Schmid for providing source codes to calculate coverage, sharpness, and contrast metrics from a specular image and Wiebke Siedentop and Annika Zentel for assisting with data collection.

TM is supported by a Sir Henry Wellcome Postdoctoral Fellowship from Wellcome Trust (218657/Z/19/Z) and a Junior Research Fellowship from Pembroke College, University of Oxford. This research was also supported by “The Adaptive Mind,” funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research, by the DFG SFB-TRR-135 “Cardinal Mechanisms of Perception” (project No. 222641018, TPs C1 and C2), and by a Marsden Fast Start Grant to KS from the Royal Society of New Zealand (project MFP-UOA2109), and European Research Council Advanced Grant Color3.0 project number 884116. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Data access: The raw experimental data, all stimulus images, and source codes to analyze the data and reproduce figures are available at https://github.com/takuma929/ColorANDGlossConstancy.

Commercial relationships: none.
Corresponding author: Takuma Morimoto.
Email: takuma.morimoto@psy.ox.ac.uk.
Address: New Radcliffe House, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6NW, UK.

References

Adams, W. J., Elder, J. H., Graf, E. W., Leyland, J., Lugtigheid, A. J., & Muryy, A. (2016). The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. *Scientific Reports, 6*, 35805.

Adams, W. J., Kucukoglu, G., Landy, M. S., & Mantiuk, R. K. (2018). Naturally glossy: Gloss perception, illumination statistics, and tone mapping. *Journal of Vision, 18*, 4, https://doi.org/10.1167/18.13.4.

Anderson, B. L., & Kim, J. (2009). Image statistics do not explain the perception of gloss and lightness. *Journal of Vision, 9*(11), 10, 1–17, https://doi.org/10.1167/9.11.10.

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

Borgonovo, E., & Plischke, E. (2016). Sensitivity analysis: A review of recent advances. *European Journal of Operational Research, 248*, 3, 869–887.

Boyadzhiev, I., Bala, K., Paris, S., & Adelson, E. (2015). Band-sifting decomposition for image-based material editing. *ACM Transactions on Graphics, 34*(5), 1–16.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436.

Brainard, D. H., Cottaris, N. P., & Radonjić, A. (2018). The perception of color and material in natural tasks. *Royal Society Interface Focus, 8*(4), 20180012.

Brainard, D. H., & Maloney, L. T. (2004). Perception of color and material properties in complex scenes. *Journal of Vision, 4*(9), ii–iv, https://doi.org/10.1167/4.9.i.

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, 9*, 1433–1448.

Chadwick, A. C., & Kentridge, R. W. (2015). The perception of gloss: A review. *Vision Research, 109*, 221–235.

Cheeseman, J. R., Ferwerda, J. A., Maile, F. J., & Fleming, R. W. (2021). Scaling and discriminability of perceived gloss. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 38*, 203–210.

de Almeida, V. M., Fiadeiro, P. T., & Nascimento, S. M. (2010). Effect of scene dimensionality on colour constancy with real three-dimensional scenes and objects. *Perception, 39*(6), 770–779.

Debevec, P. (1998). Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *SIGGRAPH98 Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 189–198). New York: Association for Computing Machinery.

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

Doerschner, K., Boyaci, H., & Maloney, L. T. (2010). Estimating the glossiness transfer function induced by illumination change and testing its transitivity. *Journal of Vision, 10*(4), 1–9, https://doi.org/10.1167/10.4.8.

Doerschner, K., Fleming, R. W., Yilmaz, O., Schrater, P. R, Hartung, B., & Kersten, D. (2011). Visual motion and the perception of surface material. *Current Biology, 21*(23), 2010–2016.

Dovencioglu, D.N., Ben-Shahar, O., Barla, P., & Doerschner, K. (2017). Specular motion and 3D shape estimation. *Journal of Vision, 17*(5), 19, https://doi.org/10.1167/17.6.3.

Dror, R. O., Willsky, A. S., & Adelson, E. H. (2004). Statistical characterization of real-world illumination. *Journal of Vision. 4*(9), 821–837. 11, https://doi.org/10.1167/4.9.11.

Ennis, R., Schiller, F., Toscani, M, & Gegenfurtner, K. R. (2018). Hyperspectral database of fruits and vegetables. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 35*, B256–B266.

Fairchild, M. D. (2013). *Color Appearance Models*. New York: John Wiley & Sons, Ltd.

Flachot, A., & Gegenfurtner, K. R. (2018). Processing of chromatic information in a deep convolutional neural network. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 35*, B334–B346.

Flachot, A., & Gegenfurtner, K. R. (2021). Color for object recognition: Hue and chroma sensitivity in the deep features of convolutional neural networks. *Vision Research, 182*, 89–100.

Fleming, R. W. (2014). Visual perception of materials and their properties. *Vision Research, 94*, 62–75.

Fleming, R. W. (2017). Material perception. *Annual Review of Vision Science, 3*, 365–388.

Fleming, R. W., Dror, R. O., & Adelson, E. H. (2003). Real-world illumination and the perception of surface reflectance properties. *Journal of Vision, 3*(5), 3, 347–368, https://doi.org/10.1167/3.5.3.

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

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. arXiv.

Geisler-Moroder, D., & Dür, A. (2010). A new ward BRDF model with bounded albedo. *Computer Graphics Forum, 29*, 4, 1391–1398.

Giesel, M., & Gegenfurtner, K. R. (2010). Color appearance of real objects varying in material, hue, and shape. *Journal of Vision, 10*(9), 10, https://doi.org/10.1167/10.9.10.

Gigilashvili, D., & Islam, A.J. (2022). The role of shape in modeling gloss, *Proceedings of the 30th Color and*

Imaging Conference. Society for Imaging Science and Technology, November 13–17, Scottsdale, USA, pp. 271–276.

Gigilashvili, D., Thomas, J., Pedersen, M., & Hardeberg, J. Y. (2021). On the appearance of objects and materials: Qualitative analysis of experimental observations, *Journal of the International Colour Association (JAIC), 27*, 26–55.

Goncalves, N. R., & Welchman, A. E. (2017). "What Not" Detectors Help the Brain See in Depth, *Current Biology, 27*, 1403–1412.

Granzier, J., Vergne, R., & Gegenfurtner, K.R. (2014). The effects of surface gloss and roughness on color constancy for real 3-D objects. *Journal of Vision, 14*(2), 16, https://doi.org/10.1167/14.2.16.

Hansmann-Roth, S., & Mamassian, P. (2017). A glossy simultaneous contrast: Conjoint measurements of gloss and lightness. *I-Perception, 8*(1), 2041669516687770.

Hedrich, M., Bloj, M., & Ruppertsberg, A. I. (2009). Color constancy improves for real 3D objects. *Journal of Vision, 9*(4), 1–16, https://doi.org/10.1167/9.4.16.

Hernández-Andrés, J., Romero, J., Nieves, J. L., & Lee, R. L. (2001). Color and spectral analysis of daylight in southern Europe. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 18*(6), 1325–1335.

Ho, Y.-X., Landy, M. S., & Maloney, L. T. (2008). Conjoint measurement of gloss and surface texture. *Psychological Science, 19*(2), 196–204.

Honson, V., Huynh-Thu, Q., Arnison, M., Monaghan, D., Isherwood, Z.J., & Kim, J. (2020). Effects of shape, roughness and gloss on the perceived reflectance of colored surfaces. *Frontiers in Psychology, 11*, 485.

Hurlbert, A. (2007). Colour constancy. *Current Biology, 17*, R906–R907.

Hurlbert, A., Cumming, B.G., & Parker, A.J. (1991). Recognition and perceptual use of specular reflections. *Investigative Ophthalmology & Visual Science, 32*, 1278–1278.

Isherwood, Z. J., Huynh-Thu, Q., Arnison, M., Monaghan, D., Toscani, M., Perry, S., . . . Kim, J. (2021). Surface properties and the perception of color. *Journal of Vision, 21*(2), 7, 1–22, https://doi.org/10.1167/jov.21.2.7.

Ishihara, S. (1973). *The series of plates designed as a test for colour-blindness*. Tokyo, Japan: Kanehara Shuppan Co. Ltd.

Jakob, W. (2010). Mitsuba: Physically based renderer. Retrieved from https://www.mitsuba-renderer.org/download.html.

Judd, D. B., MacAdam, D. L., Wyszecki, G., Budde, H. W., Condit, H. R., Henderson, S. T., . . . Simonds, J. L. (1964). Spectral distribution of typical daylight as a function of correlated color temperature. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 54*(8), 1031–1040.

Kim, J., & Anderson, B. L. (2010). Image statistics and the perception of surface gloss and lightness. *Journal of Vision, 10*(9):3, 1–17, https://doi.org/10.1167/10.9.3.

Kim, J., Marlow, P., & Anderson, B. L. (2012). The dark side of gloss. *Nature Neuroscience, 15*, 1590–1595, https://doi.org/10.1038/nn.3221.

Lagunas, M., Serrano, A., Gutierrez, D., & Masia, B. (2021). The joint role of geometry and illumination on material recognition. *Journal of Vision, 21*(2):2, 1–18, https://doi.org/10.1167/jov.21.2.2.

Landy, M. S. (2007). Visual perception—A gloss on surface properties. *Nature, 447*(7141), 158–159.

Lee, R. J., & Smithson, H. E. (2016). Low levels of specularity support operational color constancy, particularly when surface and illumination geometry can be inferred. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 33*, A306–A318.

Lee, R. J., & Smithson, H. E. (2017). Motion of glossy objects does not promote separation of lighting and surface colour. *Royal Society. Open Science, 4*, 171290.4171290171290.

Liao, C., Sawayama, M., & Xiao, B. (2023). Unsupervised learning reveals interpretable latent representations for translucency perception. *Plos Computational Biology, 19*(2):e1010878.

Marlow, P. J., & Anderson, B. L. (2013). Generative constraints on image cues for perceived gloss. *Journal of Vision, 13*, 2, https://doi.org/10.1167/13.14.2.

Marlow, P. J., Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology, 22*(20), 1909–1913.

Milojevic, Z., Ennis, R., Toscani, M., & Gegenfurtner, K.R. (2018). Categorizing natural color distributions. *Vision Research, 151*, 18–30.

Mizokami, Y. (2019). Three-dimensional stimuli and environment for studies of color constancy. *Current Opinion in Behavioral Sciences, 30*, 217–222.

Montavon, G., Samek, W., & Müller, K. (2018). Methods for interpreting and understanding deep neural networks, *Digital Signal Processing, 73*, 1–15.

Morgenstern, Y., Murray, R. F., & Harris, L. R. (2011). The human visual system's assumption that light comes from above is weak. *Proceedings of the*

National Academy of Sciences of the United States of America, 108(30), 12551–12553.

Morimoto, T., Mizokami, Y., Yaguchi, H., & Buck, S. L. (2017). Color constancy in two-dimensional and three-dimensional scenes: Effects of viewing methods and surface texture. *I-Perception, 8*(6), 2041669517743522.

Morimoto, T., Zhang, C., Fukuda, K., & Uchikawa, K. (2022). Spectral measurement of daylights and surface properties of natural objects in Japan. *Optics Express, 30*, 3183–3204.

Motoyoshi, I., & Matoba, H. (2012). Variability in constancy of the perceived surface reflectance across different illumination statistics. *Vision Research, 53*(1), 30–39.

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature, 447*, 206–209.

Nagai, T., Kaneko, S., Kawashima, Y., & Yamauchi, Y. (2017). Do specular highlights and the daylight locus act as cues for estimating illumination color from a single object? *Optical Review, 24*, 47–61.

Nicodemus, F. E., Richmond, J.C., Hsia, J.J., Ginsburg, I.W., & Limperis, T. (1997). Geometrical considerations and nomenclature for reflectance. Gaithersburg, MD: National Bureau of Standards, NBS monograph 160.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology, 10*(4), e1003553.

Nishida, S. (2019). Image statistics for material perception. *Current Opinion in Behavioral Sciences, 30*. 94–99.

Nishida, S., & Shinya, M. (1998). Use of image-based information in judgments of surface-reflectance properties. *Journal of the Optical Society of America A, 15*(12), 2951–2965.

Obein, G., Knoblauch, K., & Viéot, F. (2004). Difference scaling of gloss: Nonlinearity, binocularity, and constancy. *Journal of Vision, 4*(9), 4, https://doi.org/10.1167/4.9.4.

Olkkonen, M., & Brainard, D. H. (2010). Perceived glossiness and lightness under real-world illumination. *Journal of Vision, 10*(9). 5, 1–19. http://www.journalofvision.org/content/10/9/5, https://doi.org/10.1167/10.9.5.

Pearce, B., Crichton, S., Mackiewicz, M., Finlayson, G. D., & Hurlbert, A. (2014). Chromatic illumination discrimination ability reveals that human colour constancy is optimised for blue daylight illuminations. *PloS One, 9*(2), e87989.

Pellacini, F., Ferwerda, J. A., & Greenberg, D. P. (2000). Toward a psychophysically-based light reflection model for image synthesis. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. New York: ACM Press/Addison-Wesley Publishing Co.; 55–64, https://doi.org/10.1145/344779.344812.

Pont, S. C., & te Pas, S. F. (2006). Material-illumination ambiguities and the perception of solid objects. *Perception, 35*, 1331–1350.

Prokott, E., & Fleming, R. W. (2022). Identifying specular highlights: Insights from deep learning. *Journal of Vision, 22*(7):6, 1–19, https://doi.org/10.1167/jov.22.7.6.

Prokott, K. E., Tamura, H., & Fleming, R. W. (2021). Gloss perception: Searching for a deep neural network that behaves like humans, *Journal of Vision, 21*(12):14, 1–20, https://doi.org/10.1167/jov.21.12.14.

Radonjic, A., & Brainard, D. H. (2016). The nature of instructional effects in color constancy. *Journal of Experimental Psychology. Human Perception and Performance, 42*(6), 847–865.

Radonjić, A., Cottaris, N. P., & Brainard, D. H. (2018). Quantifying how humans trade off color and material in object identification. *Proceedings of Electronic Imaging 2018*, Burlingame, CA.

Radonjić, A., Cottaris, N. P., & Brainard, D. H. (2019). The relative contribution of color and material in object selection. *PLoS Computational Biology, 15*(4), e1006950.

Ramachandran, V. S. (1988). Perception of shape from shading. *Nature, 331*(6152), 163–166.

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

Rideaux, R., & Welchman, A. E. (2021). Exploring and explaining properties of motion processing in biological brains using a neural network. *Journal of Vision, 21*(2):11, 1–17, https://doi.org/10.1167/jov.21.2.11.

Sawayama, M., Dobashi, Y., Okabe, M., Hosokawasa, K., Koumura, T., Saarela, T., . . . Nishida, S. (2022). Visual discrimination of optical material properties: A large-scale study. *Journal of Vision, 22*(2):17, 1–24, https://doi.org/10.1167/jov.22.2.17.

Sawayama, M., & Nishida, S. Y. (2018). Material and shape perception based on two types of intensity gradient information. *PLoS Computational Biology, 14*(4), e1006061.

Scheller Lichtenauer, M., Schuetz, P., & Zolliker, P. (2013). Interaction improves perception of gloss. *Journal of Vision, 13*(14), 14, 1–13, https://doi.org/10.1167/13.14.14.

Schiller, F., Valsecchi, M., & Gegenfurtner, K. R. (2018). An evaluation of different measures of color saturation. *Vision Research, 151*, 117–134.

Schmid, A. C.,, Barla, P., & Doerschner, K. (2021). Material category of visual objects computed from specular image structure. bioRxiv 2019.12.31.892083, https://doi.org/10.1101/2019.12.31.892083.

Sharan, L., Li, Y., Motoyoshi, I., Nishida, S., & Adelson, E. H. (2008). Image statistics for surface reflectance perception. *Journal of the Optical Society of America A, 25*, 846–865.

Smits, B. (1999). An RGB-to-spectrum conversion for reflectances. *Journal of Graphics Tools, 4*, 11–22.

Smithson, H., & Zaidi, Q. (2004). Colour constancy in context: Roles for local adaptation and levels of reference. *Journal of Vision, 4*, 693–710, https://doi.org/10.1167/4.9.3.

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

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

Storrs, K.R., Anderson, B.L., & Fleming, R.W. (2021). Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour, 5***,** 1402–1417.

Tamura, H., Prokott, K. E., & Fleming, R. W. (2022). Distinguishing mirror from glass: A "big data" approach to material perception, *Journal of Vision, 22*(4), 4, 1–22, https://doi.org/10.1167/jov.22.4.4.

Toscani, M., Valsecchi, M., & Gegenfurtner, K.R. (2013a). Optimal sampling of visual information for lightness judgments. *Proceeding of the National Academy of Sciences of the United States of America, 110*(27), 11163–11168.

Toscani, M., Valsecchi, M., & Gegenfurtner, K.R. (2013b). Selection of visual information for lightness judgements by eye movements. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 368*, 20130056.

Vangorp, P., Laurijssen, J., & Dutre, P. (2007). The influence of shape on the perception of material reflectance. *ACM Transactions on Graphics, 26*(3), 77, 1–9.

Vu, C. T., Phan, T. D., & Chandler, D. M. (2012). S3: A spectral and spatial measure of local perceived sharpness in natural images. *IEEE Transactions on Image Processing, 21*(3), 934–945.

Ward, G. J. (1992). Measuring and modeling anisotropic reflection. In *Proceedings of the 19th annual conference on Computer graphics and interactive techniques* (pp. 265–272).

Wedge-Roberts, R., Aston, S., Beierholm, U., Kentridge, R., Hurlbert, A., Nardini, M., . . . Olkkonen, M. (2020). Specular highlights improve color constancy when other cues are weakened. *Journal of Vision, 20*(12):4, 1–22, https://doi.org/10.1167/jov.20.12.4.

Weiss, D, Witzel, C., & Gegenfurtner, K. R. (2017). Determinants of colour constancy and the blue bias. *i-Perception, 8*(6), 2041669517739635.

Wendt, G., Faul, F., Ekroll, V., & Mausfeld, R. (2010). Disparity, motion, and color information improve gloss constancy performance. *Journal of Vision, 10*(9), 7, https://doi.org/10.1167/10.9.7.

Wiebel, C., Toscani, M., & Gegenfurtner, K.R. (2015). Statistical correlates of perceived gloss in natural images. *Vision Research. 115B*, 175–187.

Witzel, C., & Gegenfurtner, K.R. (2018). Color perception: Objects, constancy, and categories. *Annual Review of Vision Science, 4*, 475–499.

Xia, L., Pont, S. C., & Heynderickx, I. (2017). Light diffuseness metric part 1: Theory. *Lighting Research & Technology, 49*(4), 411–427.

Xiao, B., & Brainard, D. H. (2008). Surface gloss and color perception of 3D objects. *Visual Neuroscience, 25*, 371–385.

Xiao, B., Hurst, B., MacIntyre, L., & Brainard, D. H. (2012). The color constancy of three-dimensional objects. *Journal of Vision, 12*(4), 6, https://doi.org/10.1167/12.4.6.

Yang, J. N., & Maloney, L. T. (2001). Illuminant cues in surface color perception: Tests of three candidate cues. *Vision Research, 41*, 2581–2600.

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

Zhang, F., de Ridder, H., Barla, P., & Pont, S. (2019). A systematic approach to testing and predicting light-material interactions. *Journal of Vision, 19*(4), 11, 1–22, https://doi.org/10.1167/19.4.11.

Zhang, F., de Ridder, H., Barla, P., & Pont, S. (2020). Effects of light map orientation and shape on the visual perception of canonical materials. *Journal of Vision, 20*(4), 13, 1–18, https://doi.org/10.1167/jov.20.4.13.

Zhang, F., de Ridder, H., & Pont, S. C. (2018). Asymmetric perceptual confounds between canonical lightings and materials. *Journal of Vision, 18*(11), 11, 1–19, https://doi.org/10.1167/18.11.11.