Vision Research 158 (2019) 173–188

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Disentangling simultaneous changes of surface and illumination

Robert Ennis[a,*], Katja Doerschner[a,b,c]

[a] Justus-Liebig-Universitaet Giessen, Department of General Psychology, Giessen, Germany
[b] Bilkent University, Ankara, Turkey
[c] National Magnetic Resonance Research Center (UMRAM), Ankara, Turkey

ARTICLE INFO

No of reviewers = 1

Keywords:
Color constancy
Surface constancy
Illumination judgment
Chromatic scene statistic
Global scene statistics
Local scene statistics

ABSTRACT

Retinally incident light is an ambiguous product of spectral distributions of light in the environment and their interactions with reflecting, absorbing, and transmitting materials. An ideal color constant observer would unravel these confounded sources of information and account for changes in each factor. Scene statistics have been proposed as a way to compensate for changes in the illumination, but few theories consider changes of 3-dimensional surfaces. Here, we investigated the visual system's capacity to deal with simultaneous changes in illumination and surfaces. Spheres were imaged with a hyperspectral camera in a white box and their colors, as well as that of the illumination were varied along "red-green" and "blue-yellow" axes. Both the original hyperspectral images and replica scenes rendered with Mitsuba were used as stimuli, including rendered scenes with Glavens (Acta Psychologica, 2009, 132, 259–266). Observers viewed sequential, random pairs of our images, with either the whole scene, only the object, or only a part of the background being present. They judged how much the illuminant and object color changed on a scale of 0–100%. Observers could extract simultaneous illumination and reflectance changes when provided with a view of the whole scene, but global scene statistics did not fully account for their behavior, while local scene statistics improved the situation. There was no effect of color axis, shape, or simulated vs. original hyperspectral images. Observers appear to be making use of various sources of local information to complete the task.

1. Introduction

In most natural viewing circumstances, our visual system is only provided with a distribution of light across the retina, from which it must deduce a number of objects and events, resulting in the perception of an illuminated scene. The proximal stimulation itself can be the result of a potentially infinite amount of physical interactions, with many combinations of surfaces and illuminants resulting in the same distribution of light on the retina. Since our natural environments are often composed of objects with relatively stable surface properties and we are able to perceive this even in the face of changing illumination conditions, then our visual system must have some reliable method for extracting the relative contributions of surfaces and illuminants to the proximal stimulus. Typically, color vision investigators have been interested in the ability with which the visual system can assign a stable color to surfaces in spite of changing illumination conditions. This process, and its output, is often known as color constancy, and it is one of a class of constancy phenomena, including processes such as shape constancy and material constancy.

The topic of color constancy has been the subject of a long line of research and techniques (Foster, 2011; Hurlbert, 2007; Smithson, 2005; Werner, 2014), including the likes of Helmholtz (1867) and Hering (1878), and a variety of mechanisms to achieve constancy have been proposed. Most of these mechanisms involve the computation of a scene statistic (Brainard, Kraft, & Longere, 2003), which is then used to infer, and correct for, illumination changes. For example, the mean color statistic (Buchsbaum, 1980) states that the average color for a given scene should be roughly equal to the color of the light source illuminating that scene. It is often accompanied by the Gray World hypothesis, which states that the average color of surfaces across scenes is constant (Brainard et al., 2003; Hurlbert, 1998). Removing the average color from the scene would result in a new scene whose colors should correspond to the stable surface properties of the objects in the scene. In other words, any change of the illuminant would be accounted for, resulting in a constant and stable perception of surface colors. Over time, this scheme has been given some different formulations and neural adaptation has been proposed as a potential mediating substrate (Burnham, Evans, & Newhall, 1957; Jameson & Hurvich, 1989; Kaiser &

* Corresponding author.
E-mail addresses: Robert.Ennis@psychol.uni-giessen.de (R. Ennis), Katja.Doerschner@psychol.uni-giessen.de (K. Doerschner).
URL: http://www.uni-giessen.de/fbz/fb06/psychologie/abt/allgemeine-psychologie/bapl (K. Doerschner).

https://doi.org/10.1016/j.visres.2019.02.004
Received 10 September 2018; Received in revised form 11 February 2019; Accepted 12 February 2019
Available online 21 March 2019

Boynton, 1996; Webster & Mollon, 1995). However, such a mechanism will run into trouble, since there are potentially many scenes with the same average color produced by several different combinations of surfaces and illuminants. Consider a white room under a red illuminant and a red room under a white illuminant, similar to a scenario initially proposed by Gilchrist et al. (1984). In such a case, the scenes can be constructed such that the average color is the same in both circumstances. Removing that shared component from these scenes would not get you any closer to the truth. Rather, local information in the shadows and interreflections would assist in determining the properties of the environment (Ruppertsberg & Bloj, 2007). The relationship between shadows and interreflections, as well as their effect on color signals, has been quantified for a few Lambertian and mirrored scenes, illuminated by uniform diffuse light sources (Koenderink & van Doorn, 1983; Langer, 1999; Langer, 2001; Moon, 1940).

To deal with these complications, it has been proposed that the visual system might extract more statistics than just the average color. For example, one could compute correlations along the axes defining the color space of choice. Golz and MacLeod (2002) have proposed that the visual system computes the correlation between the logarithm of the luminance values and the logarithm of the "red-green" values in a given scene (as defined in a LMS space (Stockman & Sharpe, 2000; Stockman, Sharpe, & Fach, 1999)), as a means for extracting the illuminant color (i.e., "the brighter the illuminant, the redder it is"). Other proposed mechanisms include using the brightest point in the scene as a measure of the illumination (Gilchrist et al., 1999; Land & McCann, 1971) or calculating the variance of hues in a scene (i.e., a lower variance in hues tends to correspond to an illuminant that is further from equal energy white in chromaticity space, since the spectral distribution of the illuminant is forced to collapse further and further around a single wavelength (Schanda, 2014)). The visual system could also potentially use a combination of these. However, all of these mechanisms have been primarily designed under the assumption that in most scenes, only the illumination changes and surfaces remain constant.

In fact, surface properties of objects can (and do) change rapidly. For example, adding milk to coffee, lighting something on fire (which also induces a concomitant change in the illumination), spilling liquids across clothing, painting a canvas, etc. In most of these cases, one would want to be able to recognize that the surface has changed and the illumination has remained constant. In the case of setting fire to something, one would also want to be capable of disentangling the simultaneous change of surface and illumination properties.

There have been studies in the color constancy literature that have investigated the capacity of observers to detect simultaneous changes in surface reflectance and the illumination. Craven and Foster (1992), Linnell and Foster (1996), Foster, Amano, and Nascimento (2000), and Foster et al. (2001) have investigated the behavior of observers when they were requested to detect if either the illumination on a number of simulated 2-dimensional colored patches had changed or if the patches themselves had changed. Through their investigations, they have found that observers can distinguish between reflectance changes and illumination changes in this task and that the process is fast (Craven & Foster, 1992) and observers improve with quicker changes of reflectance (Linnell & Foster, 1996). The transient signals involved have also been investigated (Foster et al., 2000) and the process acts as a global and pre-attentive mechanism (Foster et al., 2001). In addition, this group of researchers has proposed that cone excitation ratios are a physical characteristic of scenes that the visual system could use to achieve color constancy while distinguishing illuminant from reflectance changes (Foster & Nascimento, 1994). Cone excitation ratios measure how much the ratio of excitation between a pair of points changes over time. If the change in mean cone excitation ratios over a region is small, over a short period of time, then this is likely due to a similarly small, natural, illumination change or no change at all. If the change in mean cone excitation ratios over a region is large enough (relative deviation in cone-excitation ratios), then this is likely due to a reflectance change over a short period of time, a change in the spatial distribution of the illuminant, or a localized change in the color of the illuminant Foster, Amano, and Nascimento (2016) and Nascimento and Foster (2001). Naturally, shadows and borders between objects of different reflectances pose problems, but one can alter the algorithm to avoid comparing points across these boundaries (Foster et al., 2016). The cone excitation ratio statistic has been tested for both 2-dimensional scenes (Foster & Nascimento, 1994; Nascimento & Foster, 2000) and 3-dimensional natural images (Foster, Amano, & Nascimento, 2006; Foster et al., 2016), for example. In the case of Foster et al. (2006), which used 3-dimensional natural scenes, simultaneous changes of both the illumination and the test surface were investigated.

While the studies discussed above investigated potential scene statistics that an observer might use to achieve color constancy, there are potentially other sources of information, many of which could be correlated. To gain a better handle on this, work that manipulated the reliability of various cues has been done with real stimuli in natural lighting conditions (i.e., actual objects in the real world; nothing was produced on a monitor). For instance, Kraft and Brainard (1999) placed observers before a small room with a MacBeth color checker and a target patch. The apparent color of the test patch could be altered by the observer, via controls that altered the light coming from a projection system focused on the target. The conditions were controlled to minimize the chances that the observers noticed the emitted light, so that they saw the target patch as a flat piece of paper, whose color could change. The authors tested the possibility that observers either used the mean color of the scene, the color of the local surround around the target patch, or the most intense region of the scene to achieve color constancy. They did this by manipulating surface reflectances, as well as the illumination on the scene, to keep the scene statistic of interest constant, while manipulating others, effectively neutralizing its utility. For instance, in the mean color condition, a change of scene reflectance was counter-balanced by a change in the illuminant to keep the mean color the same, while many other statistics could naturally vary. In all of their conditions, the authors found that observers could maintain some degree of color constancy, although observers were best in a control condition that tested the traditional color setup, where all cues were present and only the illumination changed. An overview of this study is also provided in Hurlbert (1999) and an account in the context of theory is found in Brainard et al. (2003).

In the field of lightness constancy, Gerhard and Maloney (2010) have shown that observers were capable of discriminating lighting changes from surface changes on a 3-dimensional articulated checkerboard-like display, where a lighting change was a shift in the position of the light and a surface change was a permutation of the albedos assigned to the checks in the display. In addition, they tested the capacity of observers to detect a change in surface albedo for a given check, when either the lighting changed or the remaining checks were permuted. In that case, observers were best at detecting isolated albedo changes when the lighting also changed. A similar task has been investigated for 2-dimensional colored Mondrians by Amano and Foster (2004), finding that observers performed almost as well with simultaneous changes in surface (i.e., permutations of patch positions) and illumination as they did for illumination changes only. The authors found that combining cone-excitation ratios with a spatial average across the whole scene could be a reliable cue for the task.

Considering the previous research, we have tested the capacity of observers to separate the relative contributions of simultaneous surface and illumination changes for simple 3-dimensional scenes. It is important to note that we have not considered whether observers can detect these changes or not, since that has already been determined. Rather, we want to see if observers can report the magnitude by which the illuminant has changed and by which the reflectance has changed, especially when both change simultaneously. We then evaluated this for restricted views of the scene, in order to manipulate the information

observers could use to perform the task, similar to the work of Boyaci, Doerschner, and Maloney (2006) and Boyaci, Doerschner, Snyder, and Maloney (2006), and compared "real" images vs. simulated (i.e., rendered) versions of those images. We also investigated if observers were better for "red-green" changes vs. "blue-yellow" changes and if there was any effect of shape complexity. Lastly, we have tested if any of the previously mentioned chromatic scene statistics can account for observer behavior in our task.

In particular, we have shown observers images of a scene containing either a tennis-table ball or a Glaven (Phillips, Egan, & Perry, 2009) in a diffusely illuminated white chamber. During our experiment, an observer would view the scene for 1 s and then a new version of the scene would be displayed for 1 s, in which both the color of the object and the color of the illumination could simultaneously change to different degrees along different axes in color space. Observers were requested to state the degree of change in both the object and the illumination, using a memorized scale as reference points for 0% change to 100% change. Observers performed this task for a view of the whole scene, a view of the object only, and a view of a patch of the background only. In addition, for the spheres, we tested observer behavior for images of the original scene and physically-based rendered versions, which were designed to be roughly equivalent to the original scenes.

Our investigations have led to the following findings: (1) observers are capable of judging the relative contribution of simultaneous changes in surface and illumination color to the overall change in a scene and (2) in our task, observer behavior on average is the same for red-green vs blue-yellow changes, real vs simulated changes, and spheres vs Glavens. In all cases, the global chromatic scene statistics that we tested were insufficient to account for observer behavior. Rather, local chromatic scene statistics accounted better for their behaviour, but more work needs to be done to see how to generalize this to other situations. Lastly, we have provided here a new paradigm for testing color vision, as well as an overall benchmark of different chromatic scene statistics on the same task, something which can be expanded upon in the future.

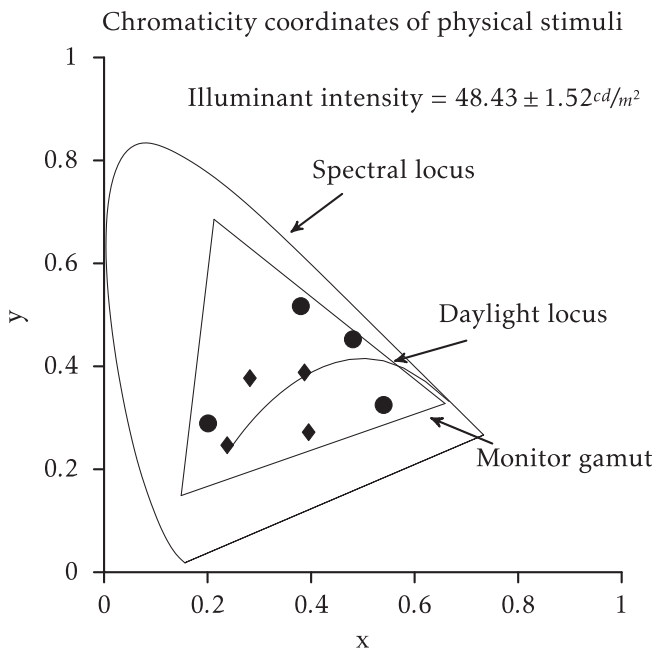

Fig. 1. Average illuminant (diamonds) chromaticity coordinates and sphere (filled circles) "chromaticity coordinates" (see text for description of their computation). Here we show the CIE1931 xyY coordinates of our physical stimuli to give an idea of the space in which we could work. The spectral locus, the daylight locus, and the gamut of our Eizo ColorEdge CG23W monitor are depicted for reference. Two of our illuminants were along the daylight locus, giving essentially blue-yellow variations, and the other two were on an axis orthogonal to this, giving red-green variations. The colors of the spheres were specified by the manufacturer, but it can be seen that two were close to being aligned with the daylight locus, giving blue-yellow variations, and the other two were essentially orthogonal to that, giving red-green variations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Materials and methods

2.1. Image acquisition and scene rendering

We wanted to test observer behavior for images of physical scenes and images of physically-based renderings of those scenes, in order to see what kinds of information about a physical environment might still be lacking in renderings. To produce such images, we first took hyperspectral images of four colored spheres under four diffuse broadband illuminants in a white chamber. The illuminants were generated by a JUST NormLicht LED box (JUST Normlicht GmbH; Weilheim/Teck, Germany) and they were constrained to be at an average intensity of 48.43 ± 1.52 cd/m^2. The spheres were standard tennis-table balls that had a diameter of 40 mm and the chamber was 25 cm × 25 cm × 27 cm.

We produced a diffuse illumination of the scene by placing two layers of photographer's Walimex Pro Diffusor diffusing paper (Walser, GmbH; Burgheim, Germany) over the surfaces of the LEDs in the LED box. An additional two layers were placed over the opening of the white chamber. This was enough to produce the percept of a "hazy, cloudy sky" illuminating the chamber (when viewed via a reflecting mirror). We confirmed the uniformity of the incident illumination by placing a Photo Research PR650 RS3 PTFE white reflectance standard (Photo Research, Inc.; Syracuse, NY, USA) at different points in the chamber and taking measurements with a Konica-Minolta CS-2000A spectrophotometer (Konica Minolta, Inc.; Tokyo, Japan). The coordinates of the spheres and the illuminants in the CIE1931 xyY chromaticity diagram can be seen in Fig. 1. Please note that the "chromaticity coordinates" of the spheres are computed from their spectral reflectance distributions and, as such, they are not proper chromaticity coordinates. The spectral reflectance distributions were obtained by measuring the spheres under a diffuse illuminant that was metameric to D65, using the white chamber already mentioned, taking the average spectral distribution and dividing that by the distribution for the illuminant. Plotting the spectral reflectances as they are in Fig. 1 is approximately the same as plotting the chromaticity coordinates for the average light reflected from the spheres if they were placed under an equal energy white. It is only meant to help visualize that the surface colors were also aligned along "red-green" and "blue-yellow" axes.

Our images were taken with a Specim VNIR HS-CL30-V8E-OEM mirror-scanning hyperspectral camera (Specim, Spectral Imaging, Ltd.; Oulu, Finland) at a spatial resolution of 57px/deg by 800px and a wavelength resolution of ~1.12 nm in the range of 376.20 nm to 821.62 nm. The device has already been described in more detail elsewhere Ennis, Schiller, Toscani, and Gegenfurtner (2018). It was confirmed that our camera could produce spectral measurements that were comparable to the CS2000-A. The raw data from our camera were calibrated and converted to radiance units ($\frac{W}{str * m^2}$) using factory measurements provided by Specim. The radiance data were then converted to CIE1931 XYZ coordinates. This calibration and conversion procedure was done by a program written in the Rust programming language (Matsakis & Klock, 2014; The Rust Programming Language, 2017).

To create a usable set of stimuli, that contained all possible variations of our illuminant colors and our sphere colors, we took images of each sphere under each of the illuminants (see Fig. 2). Linear combinations of the chromaticity coordinates of these images were used to produce variations of reflectance and illumination changes.

In the experiment, we showed observers the original hyperspectral images, as well as rendered versions of those scenes. To produce our

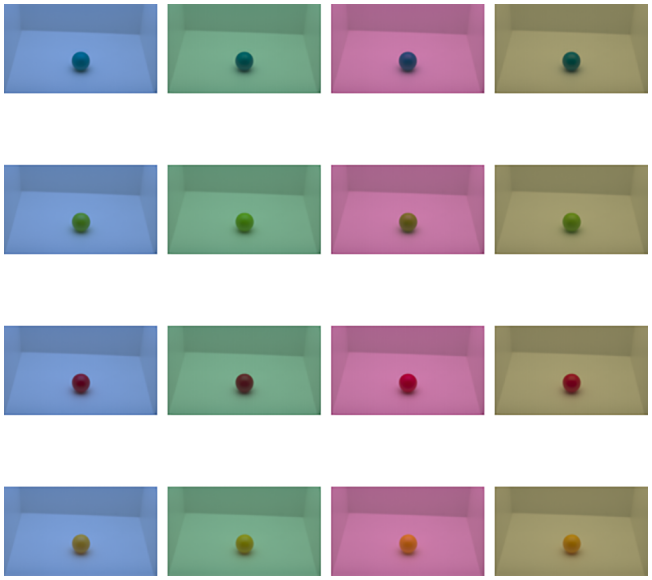

Fig. 2. Original hyperspectral images (not original size; resized to fit in the grid here). Shown here are the 16 base images used to construct variations of surface reflectance and illumination color for the sphere images. Reflectance varies across rows (blue, green, red, yellow) and illumination varies across columns (blue, green, red, yellow). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

physically-based renderings, we made use of the open-source Mitsuba renderer (v0.5.0) (Jakob, 2017). When built from its source code, one can configure this rendering system to accept spectral data of arbitrary resolution and wavelength range. We built and ran Mitsuba on an Ubuntu 15.04 LTS system (Canonical Ltd.; London, United Kingdom). We configured ours to work in the range of 360 nm to 830 nm at 47 equally spaced wavelength bands. We then built 3-D scenes in Blender (v2.76) that were equivalent in scale to the physical scenes that we took images of. These scenes were used as the base models for rendering, and the colors and positioning of the objects were set to mimic the original physical scenes. For example, a large plane was placed above a simulated chamber to act as a light emitting source, and a diffusive plane was placed just above the chamber. In addition, the walls of the scene were set to have the average spectral reflectance distributions of the walls in the original, physical scenes, as extracted from hyperspectral images taken under a diffuse illuminant metameric to D65. We also computed the average surface reflectances of our four spheres from hyperspectral images taken under the D65 illuminant, as well as the average illumination spectra. In all types of scenes, the object was rendered with the same BRDF: Mitsuba's default rough plastic BRDF, since the tennis-table balls were made of a slightly rough plastic and had a slight specular component, as judged by eye. The walls were rendered with Mitsuba's default diffuse (i.e., Lambertian) BRDF.

We made two types of rendered scenes: those with spheres (Fig. 3) and those with Glavens (Fig. 4). A Glaven is a three-dimensional deformation of a sphere formed by applying a smoothly varying noise signal to the radius at each point on the sphere's surface. One can characterize the noise signal in terms of frequency components and/or "complexity". Glavens are useful, since they have been previously studied in cross-modal vision-haptics experiments. It is known what 1 JND means for these shapes in both the visual and haptic dimensions and the digital files are freely available online (Phillips, Casella, & Egan, 2016). There were no "real" or "original" scenes for the Glavens, only simulated versions. Regardless, the spectral measurements mentioned above were applied to the rendered sphere/Glaven, as well as the rendered illumination source, to produce 16 rendered images of each sphere/Glaven under each illuminant, just as we had for the original, physical scene. A comparison of an original physical scene and a rendered scene

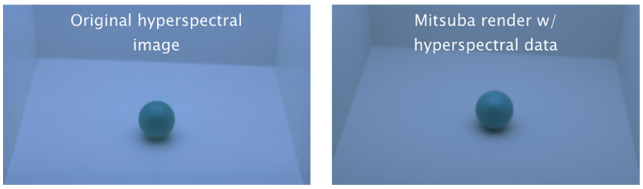

Fig. 3. Comparison of the original hyperspectral image and the physically-based rendering imitation for a given sphere and illuminant color combination. While there are some discrepancies due to the process of extracting reflectance distributions from the hyperspectral image, as well as the inherent approximations of the rendering software, one notices a striking similarity between the two images. In fact, some naive observers thought that the rendering was the "real" image.

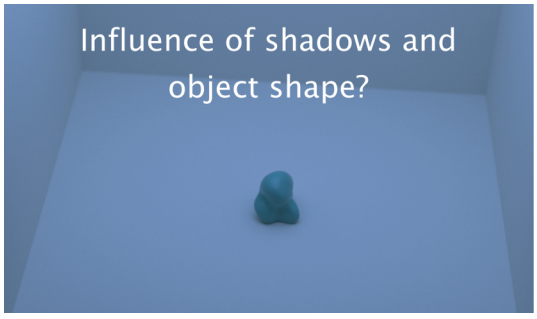

Fig. 4. An example of a rendered Glaven under the same conditions as those shown in Fig. 3. The shape is more complex than the sphere and there is a different pattern of shadows and interreflections on its surface.

based on the extracted measurements is provided in Fig. 3.

All of our final images were down-sampled to 892 px by 520 px during the actual experiment. Also, all of the images were retained in a CIE1931 XYZ representation at each pixel, rather than a linear RGB representation, to allow for accurate reproduction of the stimuli across different monitors (see "Monitors and experimental software" below).

2.2. Monitors and experimental software

For the spheres experiment, stimuli were displayed on a 10-bit EIZO ColorEdge CG23W monitor (EIZO Corporation; Hakusan, Japan) via OpenGL v3.3 using custom made software written in the Rust programming environment (Matsakis & Klock, 2014; The Rust Programming Language, 2017). Our custom made software was written to mimic the popular Processing creative programming environment (Processing Foundation). The computer that was connected to the EIZO was a Dell Precision T3610, running Microsoft Windows 7 SP1 (64-bit) (Microsoft Corporation; Redmond, Wash., USA) with an Nvidia Quadro K620 graphics card (Nvidia Corporation; Santa Clara, Cali., USA). For the Glavens experiment, stimuli were displayed on a SONY PVM2541-A OLED (Sony Corporation; Tokyo, Japan) via Psychtoolbox (v3.0.12) (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997) using the MATLAB environment, v2015a (Mathworks, Inc.; Natick, Mass., USA). The computer that was connected to the SONY OLED was a Dell Precision T3610 (Dell, Inc.; Round Rock, Texas, USA), running Microsoft Windows 7 SP1 (64-bit) with an AMD FirePro V4900 graphics card (Advanced Micro Devices, Inc.; Santa Clara, Cali., USA). Both monitors were calibrated using a Konica-Minolta CS2000-A via standard procedures documented elsewhere (Hansen & Gegenfurtner, 2013; Zaidi & Halevy, 1993). In particular, the calibrations were used to ensure that our stimuli could be accurately reproduced by the gamuts of the monitors and to calculate LMS cone excitations (Stockman & Sharpe, 2000; Stockman et al., 1999) and MacLeod-Boynton-Derrington-Krauskopf-Lennie (MB-DKL) (Derrington, Krauskopf, & Lennie, 1984; MacLeod & Boynton, 1979) representations of our stimuli, which are

detailed further in “Image statistics” below. The properties of the SONY OLED monitor in the CIE1931 xyY colorspace were as follows: red phosphor (x: 0.6751, y: 0.3224, Y: 44.51), green phosphor (x: 0.1941, y: 0.7253, Y: 102.99), and blue phosphor (x: 0.1415, y: 0.0511, Y: 11.1). The properties of the Eizo monitor in the CIE1931 xyY colorspace were as follows: red phosphor (x: 0.6588, y: 0.3277, Y: 35.98), green phosphor (x: 0.2123, y: 0.6856, Y: 66.13), and blue phosphor (x: 0.149, y: 0.0678, Y: 7.93).

Because all of our base images were stored in a CIE1931 XYZ representation and because our images were within the gamuts of both monitors, we could produce the same colors on both monitors without loss of accuracy by using the calibration data. For any given monitor, an XYZ↔(linear) RGB matrix can be created through the following procedure, where (X_R, Y_R, Z_R), (X_G, Y_G, Z_G), and (X_B, Y_B, Z_B) are the CIE1931 XYZ coordinates for each of the R, G, and B primaries for the monitor:

$$XYZ2RGB = \begin{bmatrix} X_R & X_G & X_B \\ Y_R & Y_G & Y_B \\ Z_R & Z_G & Z_B \end{bmatrix}^{-1} \quad (1)$$

2.3. Experimental procedure

In total, we ran two experiments. One experiment was finished in a single session and sessions lasted approximately one hour. In the first experiment, observers saw the scene containing a sphere and in the second, observers saw the scene containing the Glaven. Aside from the change of object, all other aspects of the scene remained the same between the two experiments. The first experiment tested real and simulated scenes, while the second experiment considered simulated scenes only. At the beginning of any experiment, observers were told that they would be “viewing images of an object in a small, white room lit by a light source from above.” It was explained that they would be shown views of either the whole scene, only the object, or only a patch of the background (see Fig. 5) and that during each trial, “either the object would change, the light would change, or both would change.” It was made clear to them that they were supposed “to state the degree of change in reflectance and illumination on scales of 0% to 100%.” After these instructions were given, observers were first shown the 16 base images on the monitor to acquaint them with the scenes and to explain to them what a reflectance change was, what an illumination change was, and what a 100% change was (i.e., since the base images were the extreme possibilities, they were the 100% changes). They were then shown a few example trials to make the instructions clear and afterwards, the actual experiment began. Each experimental session started with 2 min of adaptation to the average color of the 16 base scenes used to generate the scenes on each trial. After the initial adaptation period, a beep indicated the beginning of the experiment and observers were again shown the 16 base images to refresh their memory of 100% changes. They could press the space bar when they felt comfortable and the first trial would start with a beep, followed by one image of a scene for 1 s and then an image of the changed scene for another 1 s. The images were randomly generated using the process described in “Stimulus generation” below. Briefly, each stimulus image was some linear combination of the 16 base images with random weights used in the linear combination. The change in these random weights across each image corresponded to the changes of illumination/surface across the two stimulus images.

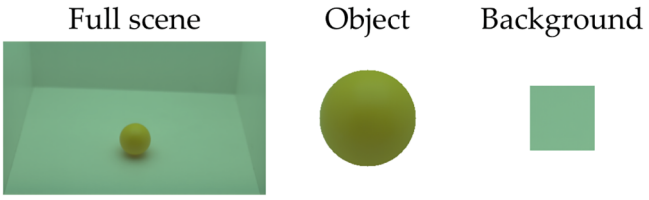

Fig. 5. The three possible views of the scene that an observer could see on any given trial. The object and the background were segmented from the whole scene view, so they were generated by the same exact process that created the whole scene view. In other words, a change in the object view could be due to a change in the object, a change in the illumination, or both. A change in the background view could also be accompanied by a change in the object, but since the object could not be seen, this had no influence. It was made clear to observers that the background view was a patch from the wall, so the response slider for the object should be set to 0% on those trials. This was done to reduce any potential confusion on the part of the observers.

While a presentation time of 1s does not allow for complete adaptation to the stimulus, this is fortunately not a problem here. While it is true that adaptation has a slow and fast component, our experiment was more about appearance than threshold level discrimination, and for appearance, most of adaptation is complete in the first 25 ms (Fairchild & Reniff, 1995; Rinner & Gegenfurtner, 2000). Due to this, we would not expect any pronounced changes in observer behavior after 10-30s, aside from perhaps a reduction in noise, since, as is detailed later in the Results, observers are already doing reasonably well. In addition, some earlier studies on color constancy have been successfully performed with presentation times as short as ours (Foster et al., 2001; Nascimento & Foster, 2000; Nascimento & Foster, 2001).

It is worth taking a moment to notice that we specifically told observers that the room was white. It could be argued that we had given the observers too much information about the stimulus and they could easily preform the task by cognitively inferring the illumination color and then using that to cognitively infer the object color, sidestepping any perceptual processes that contribute to color constancy. We argue that this is not the case. First, unfortunately, it is not really possible to have observers do the task and not tell them that the walls are white. Given our paradigm, observers must be shown examples of 100% color changes and be told, “this is a 100% change from a blue illuminant to a yellow illuminant’, for example. From this alone, one can easily infer that the room is white, or at least, that it is close to neutral and doesn’t change color throughout the experiment. Also, some observers asked if the walls could change color during the experiment. It had to be made clear to them that the walls do not change color. If the walls did change color or if an observer at least thought that the walls could change color, then the task would probably be too difficult (although that would need to be tested). Considering this, we decided to reduce confusion from the start by just telling observers that the walls were white. Aside from all of this, while our task and results could potentially derive from “cognitive color constancy”, we think it would be valuable to further distinguish between perceptual and cognitive color constancy and even to see if observers use statistics classically ascribed to one domain to solve a task in a very similar domain. Lastly, when one looks at the stimulus, especially for more perceptible magnitudes of color change, it is clearly seen as a perceptual illumination change and a perceptual reflectance change and responses can be given quite rapidly, without slower cognitive processes.

After the stimulus presentation finished, two response sliders were shown: one with an “L” (for “Light”) above it for the observer to indicate their judgment of the illumination change and the other with a small sphere above it for their judgment of the object’s reflectance change. Observers used the left mouse button to set the sliders and pressed the right mouse button to proceed to the next trial. Before the start of the next trial, there was another 1 s of adaptation to the mean color of the 16 base images. For the scene with the spheres, the experiment continued this way until 25 trials were performed for each pair of surface/illuminant color change directions (e.g., red-green change for surface and blue-yellow change for illuminant) and for each of the 3 viewing conditions. In total, this resulted in 300 trials per condition (real/simulated). For the Glavens, the number of trials for each pair of surface/illuminant color change directions was reduced to 20, resulting in 240 trails for the simulated Glaven scenes. Every 30

trials, the 16 base images were displayed again to refresh the memory of observers about 100% changes in surface and illumination. When they felt that their memory was sufficiently refreshed, they could press the space bar and the experimental trials proceeded.

In addition, before each image for a trial was cleared from the screen, we saved a screenshot using OpenGL's glReadPixels() command in order to analyze the actual images that were shown to observers, after the graphics card had carried out all of its automatic processing.

2.4. Stimulus generation

The stimuli presented on each trial were pairs of scenes, between which either the surface, the illuminant, or both changed. For a given trial, one of the four color directions was chosen for the surface and one was chosen for the illuminant and the corresponding pairs of base images were linearly combined with random weights to produce two stimulus images. For example, if a red-green surface change and a blue-yellow illumination change were chosen, then the base images of the red object under the yellow illuminant (I_{RY}), the red object under the blue illuminant (I_{RB}), the green object under the yellow illuminant (I_{GY}), and the green object under the blue illuminant (I_{GB}) would be taken. Next, a random weight for the illumination change (W_I) and a random weight for the surface change (W_S) would be produced and used to linearly combine the CIE1931 XYZ values of these images at each pixel in the following manner:

$$W_I \cdot (W_S \cdot I_{GB} + (1 - W_S) \cdot I_{RB}) + (1 - W_I) \cdot (W_S \cdot I_{GY} + (1 - W_S) \cdot I_{RY}) \quad (2)$$

This formula produces one image that is equivalent to linearly combining the illuminants and linearly combining the surface reflectances in the scene and can be applied to any representation of the image that is a linear transformation of the XYZ values (e.g., the spectra or linear RGB values). It is important to note that the process is intended for scenes composed solely of matte-like objects and across which only the properties of objects change, not their positions or shapes. In our case, the tennis table balls were made of slightly rough plastic with a slight specular component and the spheres and Glavens in the simulated scenes were rendered with the default rough plastic BRDF from Mitsuba. However, the amount of specular reflection was considered negligible, especially when compared to the amount that is typically found on glossy surfaces, and we used the image-based stimulus generation that was just discussed for our experiment. Regardless, even if the specularity is strong enough to preclude our image generation procedure, this actually still serves as a sufficient test of the statistics, since as shown below, observers can still reliably perform the task. The point is not to find the stimuli that best satisfy the statistics, but to find statistics that best match observers in all conditions.

2.5. Observers

5 observers participated in the experiment with the sphere and a separate 5 observers participated in the experiment with the Glavens. All observers were in the age range of 20–30 years old, so yellowing of the macular pigment cannot be considered a significant contribution to our results. They were naïve to the purpose of the experiment. All observers had normal or corrected to normal visual acuity. Observers were paid for their participation in the experiments. All observers gave written informed consent in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans. The experiments were approved by the local ethics committee LEK 2015-0021.

2.6. Analysis

2.6.1. Observer performance

To assess whether observers were even capable of performing the task, we first focused on responses for views of the whole scene and compared the illumination judgments that each observer made with the physical illumination change that happened on each trial, as measured by the change in the weight, W_I. The same was done for surface change judgments and the change in W_S across both images. Essentially, if observers can perform the task to some degree or another, then one should see a monotonically increasing trend: as the illumination change becomes larger, the observer's response in the illumination slider should increase, and the same for a surface change. If observers can perform the task reasonably well, then they should also at least register a 0% physical illumination change as much lower than a 100% change. If observers perform the task perfectly, then we should see a perfect linear trend. The same applies for the surface changes. We have considered the following comparisons: real vs. simulated scenes, scenes with spheres vs. those with Glavens, and "red-green" vs. "blue-yellow" changes. In particular, the deformed shape of the Glavens introduces additional shadows and lighting on its body, known as interreflections or mutual illumination, which could potentially assist observers in judging the illuminant when given a view of the object only (Bloj, Kersten, & Hurlbert, 1999; Funt & Drew, 1993; Funt, Drew, & Ho, 1991; Gilchrist et al., 1984; Ruppertsberg & Bloj, 2007). We always considered the three views of the scene separately (see Fig. 5). All data analysis was done in R (v3.4.2). Without getting too far ahead of ourselves, it was found that observers can perform the task, even though they say that they find it difficult at first.

At the end of each session, some trials had to be rejected. Essentially, if a slider was not changed during the response stage of a trial, then no data was saved for that slider. On some trials, the random position of one or the other slider was presumably satisfactory to observers, so they had not changed it, but this rendered the trial essentially useless, since the data for both sliders was necessary to make complete sense of it. We discarded these trails, but this did not seriously impact data analysis, since this only occurred for 3.87% of trials on average for the hyperspectral images of the spheres, for 4.5% of trials on average for the rendered images of the spheres, and for 6.3% of trials on average for the Glavens.

2.7. Image statistics

Aside from seeing if observers can perform the task, we wanted to determine if any of the commonly used chromatic scene statistics found in the color constancy literature could explain observer behavior. We performed analysis on the randomly produced images that were saved from each trial (see "Experimental procedure" above). The image analysis software was written in the Rust programming language (Matsakis & Klock, 2014; The Rust Programming Language, 2017). For each pair of test images, we examined the change in the following chromatic scene statistics, where *LMS* and *DKL* are images in the respective colorspaces (i.e., the LMS cone activation space and the second-stage cone opponent MB-DKL space), $\{LD, RG, YV\}$ are the cardinal axes of the MB-DKL space, n is the number of pixels per image, and p is a given pixel. In our case, the implementations of the statistics are inspired by previous reports and our formulations were as follows:

Luminance-redness (Pearson) correlation (Golz & MacLeod, 2002):

$$LRC(LMS) = r(\log_{10}(L + M), \log_{10}(\frac{S}{L + M})). \quad (3)$$

Mean cone excitation ratios (implicitly compares both images) (Nascimento & Foster, 2001):

$$CER(LMS) = \langle \frac{\|\boldsymbol{r} - \boldsymbol{r}'\|}{\frac{\|\boldsymbol{r}\| + \|\boldsymbol{r}'\|}{2}} \rangle, \quad (4)$$

where $\boldsymbol{r}$ and $\boldsymbol{r}'$ are $[L, M, S]$ excitation ratios for the same two randomly chosen points in the first and second LMS image, respectively. The sampling procedure was the same as that in Foster et al. (2016). Note

that the denominator in our formulation of the mean cone excitation ratios is slightly different from that in Foster et al. (2016) in order to keep the final value of the mean cone excitation ratio a scalar quantity.

White patch (Land & McCann, 1971):

$$WP(DKL) = DKL_{[LD,RG,YV]}[p_m], \tag{5}$$

where p_m satisfies $\max_{p \in DKL} DKL_{LD}[p]$.

Average color (Buchsbaum, 1980):

$$AC(DKL) = [\langle DKL_{LD}\rangle, \langle DKL_{RG}\rangle, \langle DKL_{YV}\rangle]. \tag{6}$$

Hue variance (Inspired by Brown & MacLeod, 1997, but not the same as their concept. Our version is based on the idea that the variance of hues should decrease as the illuminant becomes more saturated.):

$$HV(DKL) = 1 - \|\frac{1}{n}\sum_{p=1}^{n} e^{i\cdot atan2(DKL_{YV}[p], DKL_{RG}[p])}\|. \tag{7}$$

Note that our version of the White patch statistic is essentially a variant of the "brightest region is most helpful for extraction of albedo/surface color" rule (Giesel & Gegenfurtner, 2010; Gilchrist et al., 1999; Toscani, Valsecchi, & Gegenfurtner, 2013).

The LMS and DKL images were formed using the calibration data of the monitors. Briefly, an $\{L, M, S\}$ triplet can be computed from an $\{R, G, B\}$ triplet in the range of [0, 1] with knowledge of the spectral distributions emitted from the three primaries when they are at their maximum intensity. Once these were obtained, we used 2° LMS cone spectral sensitivity functions Stockman et al. (1999) and Stockman and Sharpe (2000) to calculate the $\{L, M, S\}$ excitations for the three primaries. Then, provided that the primaries do not change in their properties as their intensity changes, one can take a given $\{R, G, B\}$ triplet and scale and sum the maximum $\{L, M, S\}$ excitations accordingly to get the total $\{L, M, S\}$ excitation. For example, if the $\{R, G, B\}$ triplet $= \{0.5, 0.2, 0.5\}$, then the total $\{L, M, S\}$ excitation $= \{0.5\cdot(L_R + L_G + L_B), 0.2\cdot(M_R + M_G + M_B), 0.5\cdot(S_R + S_G + S_B)\}$. The conversion from an $\{R, G, B\}$ triplet to the DKL space is covered in Zaidi and Halevy (1993) and Hansen and Gegenfurtner (2013) and involves finding the combinations of $\{R, G, B\}$ values that independently activate the cone-opponent retinal ganglion cells (Derrington et al., 1984).

Once the image statistics were computed, we computed the change in them between the pairs of images shown on each trial. We evaluated whether the changes in the statistic values correlated with the surface or the illumination responses of observers. This was first computed globally for each of the three views of the scene.

Global chromatic scene statistics have been previously proposed as a source of information for achieving color constancy. Since our experiment is related to color constancy, in that observers must extract information about the illumination and surfaces to perform the task, we wanted to see if any of the more common chromatic scene statistics correlate with observer behavior. In particular, we took each pair of images that an observer saw on each trial and computed our statistics for the whole image at once. We then simply took the difference between the values computed for each pair of images from each trial, except for the mean cone excitation ratios, which already implicitly compare the difference between two images. Also, since both the White patch and Average color statistics return vectors, the vector response for the second image (i.e., the color estimate) was subtracted from that for the first image and then the length of this vector was saved as the final value for the statistic. This process of correlating the outputs of the statistics with the observers' responses is a simple first approximation to see how informative each statistic is for performing our task.

We also computed the statistics locally. This was only done for the images with a view of the whole scene. First, the object and a 101 px by 101 px patch from the background were segmented from each image. The background patch was from the lower right corner of the room, so that interreflections at the junction would be included. Next, computations of the statistics on the object were taken as measurements of the surface change and computations of the statistics on the background patch were taken as measurements of the illumination change. In cases where it was possible, the estimated color of the illuminant given by a statistic was factored out of the image before applying said statistic to the segmented object. The color was factored out by taking each pixel in the segmented object image, going through each color channel at that pixel, and dividing the value in the channel by the illumination estimate for that channel. This was done so that local computations of the object's color more closely matched what is to be expected in a color constancy process. This was done for the White patch and Average color statistics. Please note that both of these statistics are computed in the DKL color space, where the axes are typically scaled to span the range of $[-1, 1]$. However, estimates of illuminant power and surface reflectance typically work on normalised scales of [0, 1], so one must first normalise the image and the illumination estimate to be in this range before simulating the color constancy procedure. This is done by dividing each DKL coordinate by 2 and adding 0.5. This can be done before or after the estimate of the illuminant color, but it must also be done for the illuminant color. Regardless, in all cases, the final values were then individually correlated with the surface and illumination judgments of the observers, respectively.

The comparison between global and local measures of our chromatic scene statistics was deliberate. In our case, both the illumination and the surface can change, which can lead the global scene statistics astray. If global scene statistics fail to explain observer behavior, then we can at least assume that observers parse the scene in some manner during our task, computing the statistics locally to obtain different sources of information about different features of a scene.

3. Results

3.1. Observers can disentangle simultaneous changes in illumination/reflectance

3.1.1. Sphere scene

We have found that observers can mostly disentangle the effects of simultaneous changes in illumination and reflectance. Fig. 6A shows the responses of observers. Physical changes in both surface color and illumination color are plotted against the responses that observers placed in the response sliders after each trial. One should keep in mind three things when viewing Fig. 6A. First, the values along the x-axis are the absolute magnitude of the change in the weights used to linearly combine our base images when generating stimuli.

Also, we randomly sampled from the space of possible illumination and surface color changes, so the points in the figure are binned at intervals of 10% (so, the first bin includes the range [0, 10), for example). In addition, during the experiment, for any given illumination change (let's say 50% for example), many possible surface color reflectance changes could have simultaneously occurred. The same is true for any surface reflectance change: there can be many trials with the same surface reflectance change, but many different illumination color changes. So, for any point in Fig. 6A, say one which shows observer responses for illumination color change, we are averaging observer illumination responses across all of the surface reflectance changes that occurred for trials containing the illumination color changes within the respective bin. Essentially, plotting the results as they are in Fig. 6A allows one to see the degree to which simultaneous changes in surface color interfere with judgments about the illumination and vice versa.

If we first consider the blue points, which are observer responses for a view of the whole scene, and only look at the first row of Fig. 6A, which contains responses to the surface reflectance change, we see that observer responses show a roughly linearly increasing trend with increases in the magnitude of surface color change. This is what one would expect if observers are capable of disentangling simultaneous changes in surface/illumination color. What is most important is that it mainly monotonically increases. It is important to keep in mind that the curves do not necessarily need to be linear to indicate that observers are

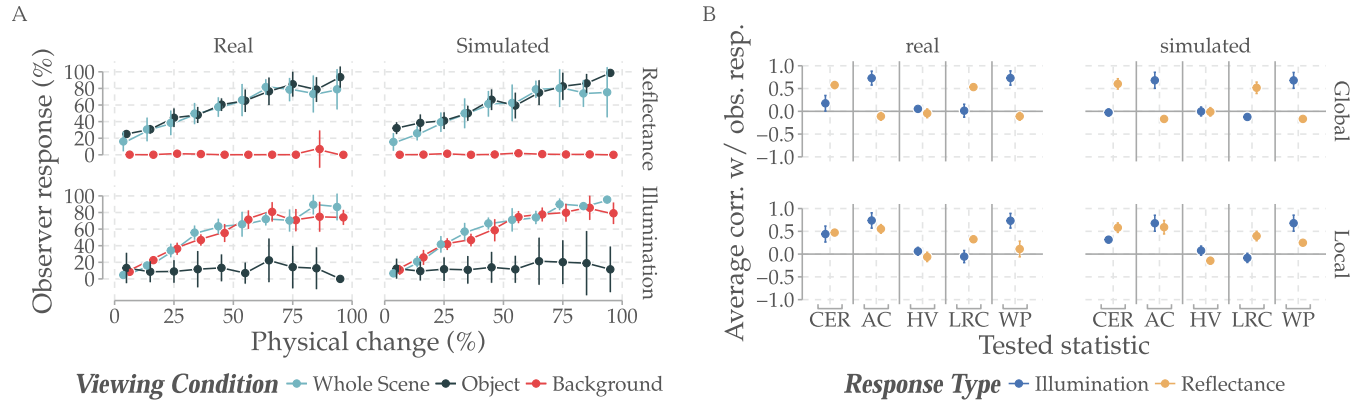

Fig. 6. Observer responses and average correlation with scene statistics for the scene with the sphere. (A) Observers responses have been partitioned into "real" & "simulated" images and "illumination" & "reflectance" judgments. On the x-axis is the physical change of either the illumination or the reflectance (as defined in the Methods) and on the y-axis is the response that the observer left in the sliders shown after each trial. In each graph, three colored sets of points are shown for the three viewing conditions: whole scene (blue), object only (black), and background patch (red). The points show the centers for the sampling distribution of the mean (i.e., the average taken over each observers average response), binned at intervals of 10% along the x-axis. The points themselves have been slightly displaced horizontally to improve visibility. Error bars show the standard error of the mean (SEM). When provided with a view of the whole scene, observers' reflectance and illumination judgments followed the magnitude of the physical reflectance and illumination changes. When viewing the sphere only, the reflectance change that contributed to the stimulus was estimated roughly correctly by observers, but any accompanying illumination change that also influenced the stimulus was inconsistently and weakly detected. When viewing the background patch only, all changes were correctly registered as illumination changes (i.e., observers essentially always set the reflectance slider to 0) and observers' illumination responses followed the actual physical change. (B) Average correlation between changes in scene statistics with observer responses. The scene statistics were computed for a view of the whole scene only. The correlations have been partitioned into "real" & "simulated" images and "global" & "local" computation of statistics. Blue symbols are for illumination responses and yellow symbols are for reflectance responses. The points are the average correlation (as computed via Fisher's z-transform) between the statistics and observer responses. The errorbars show the standard deviation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

performing the task "well". The reason is that an observer's measure of illumination/surface color change cannot be said to be a direct measure of either the physics involved or of activation at some level of the visual system. While it is related to these processes, the connection is not direct and is not necessarily linear, so we only require that the trends increase monotonically. In addition, the results show that observers' memory for the space of possibilities is not interfering with the task. If we see a perfect linear trend, then this indicates that observers are doing excellent, but even some deviation from linearity would show that they are capable of accepting and memorizing arbitrary scales of color change for our scenes.

One should also notice that observers do have some variability in their responses. This variability could be due to potentially different reasons, such as criterion effects or the inability to precisely locate the desired position on the response sliders, among other potential influences, such as the shape of psychometric functions for detecting changes in the regions of color space that we tested. However, the point still stands that in spite of all of this and in the face of simultaneous changes in illumination/reflectance, observers still have the ability to determine the relative contribution of both sources to changes in a scene.

In panel B of Fig. 6, one will find the correlations of the various scene statistics with observer responses, split up according to "real" hyperspectral images, simulated renderings, and global vs. local statistics, and plotted according to whether the statistic was correlated with the reflectance or illumination responses. Focusing first on the top row, with the global scene statistics, we find that they are not fully capable of explaining observer behavior. Sometimes one statistic is able to account well for illumination responses, such as the AC or WP statistics, and sometimes for the reflectance responses, such as the CER or LRC statistics, but when computed globally, none can account for responses to both quantities. In fact, one should not find the high correlation of the AC statistic with illumination responses to be particularly interesting. It is already known that the AC statistic does not account for a few color constancy results and there are situations where it can be misleading if one depended on it alone (Maloney, 1999). Rather, the AC statistic is doing well at predicting observers' illumination responses, because for our stimuli, it essentially reproduces the actual physical manipulation that we did. The AC statistic, for example, computes the average color of whatever it is looking at, so when it is applied to our whole image, it essentially extracts the illuminant color, since our images are dominated by the background and our background was a box composed of white surfaces. Because of this, any change in the AC statistic is essentially equivalent to the change in the illuminant color, but this is just reproducing the stimulus, which we already know correlates well with observer responses, as we have seen in panel A. Similarly, the WP statistic gives the color of the most luminant part of the scene, which in our stimuli, will always be on the white reflecting background. Its estimate of the illuminant color will be highly correlated with the AC statistic in our case, so it is also essentially reproducing the stimulus. While reproducing the stimulus has more or less been the holy grail of color constancy, the AC statistic at least already has some known deficiencies that make it an unlikely candidate for a good color constancy scene statistic (Gilchrist et al., 1984; Golz & MacLeod, 2002; Maloney, 1999). The WP statistic on the other hand is known to be diagnostic of the illuminant color in a variety of circumstances and has been linked to observer behaviour through eye movement experiments (Toscani et al., 2013).

We also find that the globally computed LRC statistic correlates decently with observers' reflectance change responses, but poorly with their illumination change responses. The LRC statistic makes use of the correlation between luminance and "redness" (i.e., excursion in the "red" direction of the DKL "red-green" axis) that has been found in a few natural scenes (Golz & MacLeod, 2002). This is to be expected, since even though the intensity of our illuminants was kept essentially constant, the spheres were not matched in this respect. In other words, for the same illumiant, the green sphere could reflect a distribution with greater luminance than the red sphere, introducing a correlation between luminance and "redness" in that region of the scene. In spite of this, the LRC is unable to explain observers' judgments of the illuminant, since the change in our illuminant does not exhibit the correlation that it seeks, so observers must be using some other source of information to complete our task. Lastly, hue variance is always doing poorly and we ignore it for the remainder of this paper.

As it stands, the global scene statistics that we tested, do not fully

account for observer behavior. Rather, the statistics generally show a poor or moderately decent correlation with observer responses (see Fig. 6B). It could also be argued that these statistics do not fully account for observer behavior because they are being computed for simple scenes with little variation in surface reflectance or complex shape. However, we would rather find a new formulation of chromatic scene statistics that explains observer behavior in our task, instead of finding a scene that is ideal for each statistic. There is also an argument against global chromatic scene statistics ever being able to account for observer behavior in our task. Since global statistics reduce the whole scene to one number, a good deal of information is lost and many scenes will produce the same value for the given statistic. The point is that because global scene statistics have historically been considered mainly for the case in which only a mostly uniform illuminant changes color, they have done fairly well at explaining observer color constancy behavior in those tasks. In our case, they fail, and it will be important to revise them to account for our task, since there are situations in the natural world that can lead to simultaneous changes of illumination and reflectance, such as when something is burning. In our case, we have considered if computing the statistics locally allows them to better parse the effects of simultaneous illumination/reflectance changes. For example, computing a change in a given statistic for a background patch could provide a clue to the illumination change and computing the same statistic for the object separately could provide a clue to the reflectance change.

In addition, one may claim that a non-linear relationship is needed to account for the leap from a statistic's output to an observer's final judgment of magnitude change. First, we tested other correlation coefficients, such as Spearman's and Kendall's tau, both of which better handle non-linear relationships, and they made no qualitative difference (i.e., the pattern of results remained the same, for both the global and local scene statistics). However, the distinction between the perceptual statistic and the link to the output of the cognitive process that becomes an observer's response is not always immediately clear, especially when the information that is most relevant to the task has not been fully clarified, as is the case here, and is outside the scope of this paper. Regardless, if we would need to include an additional non-linear process to account for observer behavior in our task, then this only provides extra support to our method of analysis, since it would make clear that the frameworks for the currently tested statistics are insufficient and more work needs to be done.

We computed the local versions of our scene statistics for a view of the whole scene by segmenting the object and a patch of the background and computing the statistics on them. These can be seen in the second row of Fig. 6B. Please remember that for the AC and WP statistics, we also used their estimate of the illuminant color to perform a rudimentary color constancy operation for obtaining the surface color by factoring it out of the image. Each of these measures were then correlated with the surface and illumination judgments of observers, respectively. We found that the performance of the CER statistic increased for estimates of the illumination change and the AC and WP statistics increased for estimates of the reflectance change. Please see above for why the AC statistic, at least, is not to be relied upon for illuminant estimation, at least in our task. However, the AC statistic is doing well at predicting observers' responses for reflectance changes. This is considered in more detail in the Discussion. The remaining statistics estimates are not improved enough by a local computation to be considered relevant for our task.

3.1.2. Glaven scene

We have found a similar pattern of results for the scene with the Glavens as we found for the spheres. In particular, observers can disambiguate simultaneous changes of illumination and reflectance changes, while global scene statistics cannot. Local scene statistics improve the situation similarly to before and in the Discussion we consider the relevance of this to previous results and the strategy that observers could be using in our task. See Fig. 7 for further details.

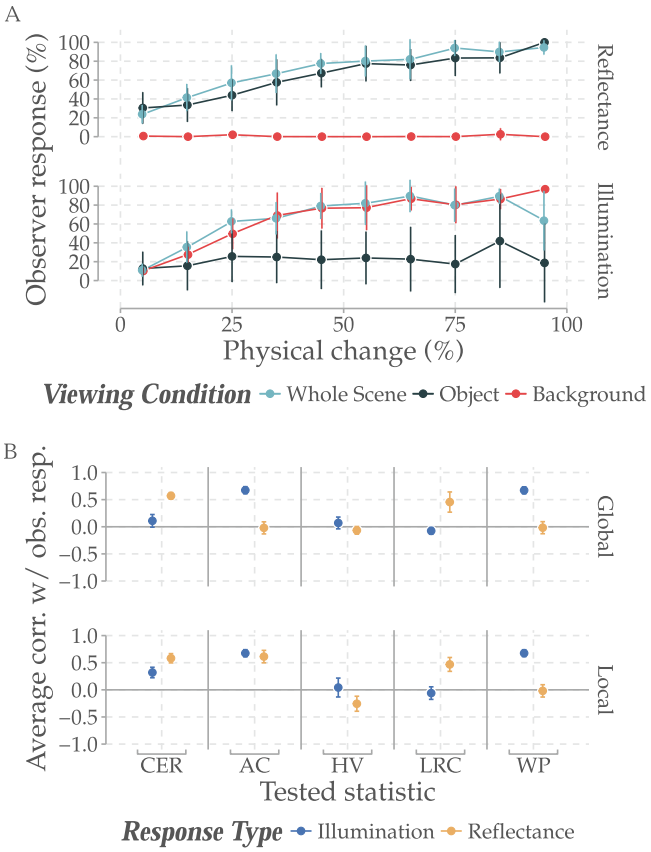

Fig. 7. Observer responses for scenes containing a Glaven. Plotting conventions follow those in Fig. 6. The difference is that now data are only shown for simulated renderings, since we did not make "real" images of scenes with Glavens.

3.2. Differences between conditions

3.2.1. No preference for color direction

It has been hypothesized in the past that observers might be more color constant for lights along the daylight locus, since this was the major locus of illuminant color change during most of our evolution (Delahunt & Brainard, 2004; Pearce, Crichton, Mackiewicz, Finlayson, & Hurlbert, 2014). The data on this are at odds with each other, though, with (Delahunt & Brainard, 2004) finding no preference for colors along the daylight locus, while (Pearce et al., 2014) found that observers are more color constant for lights along the daylight locus. Considering this, we have checked whether our observers were better for blue-yellow changes vs. red-green changes, both for the illuminants, as well as the surfaces. In particular, to make the comparisons in this and later sections, we fit Generalized Additive Models (GAMs) to the data and tested if fitting them separately for red-green vs. blue-yellow changes provided a better fit than just fitting them for the pooled data. If fitting them separately is better, then this is evidence that observers act differently for red-green changes as compared to blue-yellow changes. Briefly, GAMs are an extension of Generalized Linear Models that can be used for this type of comparison, since we only wish to compare trends and do not have a predefined function that we wish to fit to our data Knoblauch and Maloney (2012). They can also handle the non-linearities in some of the data, while essentially reducing to lines when the data show a clear linear trend. Essentially, GAMs fit smooth functions, such as splines, to the data and reduce the chances of over-fitting by using a optimization penalty that is based on the curvature of the fit. One can specify that the smooth functions be fit separately for each level of a given factor (e.g., color axis) or that one smooth function be fit for all of the data.

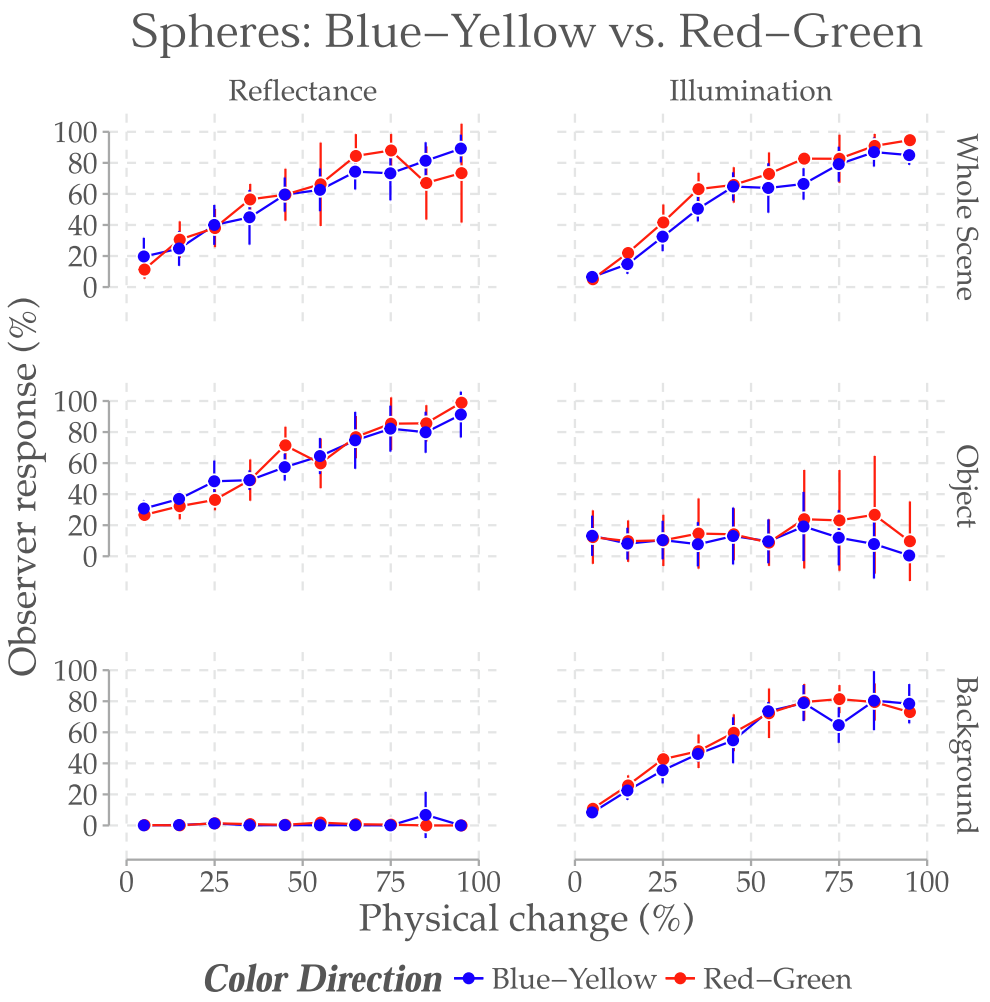

Fig. 8. Sphere scene: Observer responses for the two response types (illumination/reflectance) and the three views (whole scene, object only, or background only) parceled according to the axis along which a color change took place. Blue-yellow changes in blue and red-green changes in red. It can be seen that observers were not acting differently for one color direction over the other. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Let us first consider Fig. 8. For the sphere scenes, one can see that regardless of whether we consider surface or illumination, performance was basically the same for red-green as it was for blue-yellow changes. The same can be seen in Fig. 9, where data for the Glaven scenes is shown. At least for our task, it seems there is no preferred color axis. More specifically, there were no significant differences between nested GAMs fit separately to the data from the red-green and the blue-yellow conditions for each combination of response type (Illumination vs. Reflectance) and scene type (Whole Scene, Object, Background), as shown in Tables 1 for the sphere and 2 for the Glaven. Please note that in these tables, sometimes the probability of the null hypothesis (i.e., that a single smooth function fits both sets of data simultaneously) being true was so high that R's ANOVA routines do not produce a p-value or F ratio, so we list them as NA, i.e., as essentially blank and irrelevant. We have also included the Bayes Factor (BF), computed according to the Schwarz criterion (Kass & Raftery, 1995), that measures the weight of evidence towards one or the other model, since a lack of significant difference between the two models can sometimes be misleading (Wagenmakers, 2007). In our case, if it is greater than 1, then the simpler model is preferred and if it is less than 1, then the more complicated model is preferred. It can be seen that in all cases, the simpler model that does not fit a separate smooth function for each color axis is preferred.

3.2.2. No difference between "real" and simulated images

Physically accurate rendering of 3-dimensional scenes is still rapidly developing, so it can still be the case that a final render will have differences from the original scene that it is based on. However, for simpler scenes and simpler materials, whose optical properties are well known, the final render usually only shows small differences, if any, from the original scene, as evidenced by the frequent use of physically-based rendering systems in architectural prototyping (Ward, 1989). Here, we have tested if the renders produced by the Mitsuba rendering system produced any measurable psychophysical differences from the hyperspectral images of the original scene. If differences exist, then the renders probably lack a key feature that assists observers in performing our task, but the data do not support this, as can been seen in Fig. 10. There was only one significant difference between nested GAMs fit separately to the data from the hyperspectral image and the simulated image conditions for each combination of response type (Illumination vs. Reflectance) and scene type (Whole Scene, Object, Background), as shown in Table 3. Again, in all but one case, the BFs prefer the simpler model.

3.2.3. Glavens do not appreciably influence observer performance

Based on the work of Ruppertsberg and Bloj (2007), we tested whether using Glavens as a test object improves observer performance relative to a sphere, since the deformed bodies of the Glavens will have more interreflections on their surfaces, as well as more shadows. This

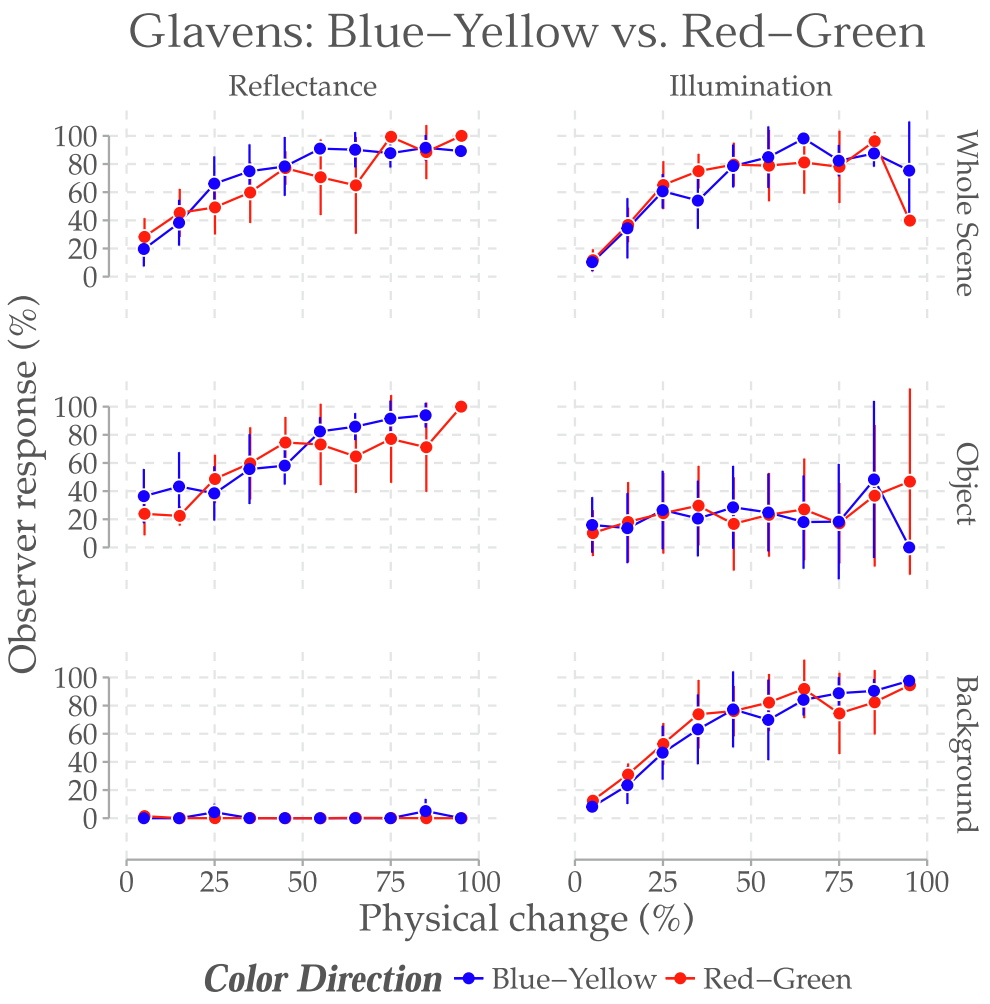

Fig. 9. Glaven scene: Observer responses for the two response types (illumination/reflectance) and the three views (whole scene, object only, or background only) parceled according to the axis along which a color change took place. Blue-yellow changes in blue and red-green changes in red. It can be seen that observers were not acting differently for one color direction over the other. The one missing point for responses to a Blue-Yellow reflectance change in the 90–100% bin of the Object only condition is due to our method for randomly selecting stimuli never having sampled in this region. It fortunately has no effect on our conclusions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Results from an ANOVA analysis comparing GAMs fit to the data shown in the six panels of Fig. 8. The models were nested, with the first fitting one smooth function to perceived magnitude of stimulus change for both color axes and the second fitting a separate smooth function for each color axis. No significant differences were found between the two models at a $p < .05$ level, indicating, in general, no differences based on the axis along which a color change took place. In addition, the BFs consistently prefer the simpler model that does not fit a separate smooth function for each color axis.

	Illumination	Surface
Whole Scene	F(81.60, 2.69) = 0.108, p = .943, BF = 108.66	F(87.05, 1.34) = NA, p = NA, BF = 15.93
Object	F(90.00, 1.00) = 0.039, p = .844, BF = 9.50	F(90.00, 1.00) = 2.910, p = .091, BF = 2.17
Background	F(78.77, 2.77) = NA, p = NA, BF = 147.28	F(88.46, 1.54) = 1.397, p = .251, BF = 6.24

Table 2
Results from an ANOVA analysis comparing GAMs fit to the data shown in the six panels of Fig. 9. The models were nested, with the first fitting one smooth function to physical magnitude of stimulus change for both color axes and the second fitting a separate smooth function for each color axis. No significant differences were found between the two models at a $p < .05$ level, indicating no differences based on the axis along which a color change took place. In addition, the BFs consistently prefer the simpler model which does not fit a separate smooth function for each color axis.

	Illumination	Surface
Whole Scene	F(77.27, 2.79) = 0.511, p = .663, BF = 66.16	F(77.22, 1.97) = 1.502, p = .229, BF = 7.26
Object	F(81.00, 1.00) = 0.467, p = .496, BF = 7.22	F(83.00, 1.00) = 0.245, p = .622, BF = 8.20
Background	F(82.15, 2.46) = 0.569, p = .604, BF = 39.22	F(77.22, 1.97) = 1.502, p = .229, BF = 7.26

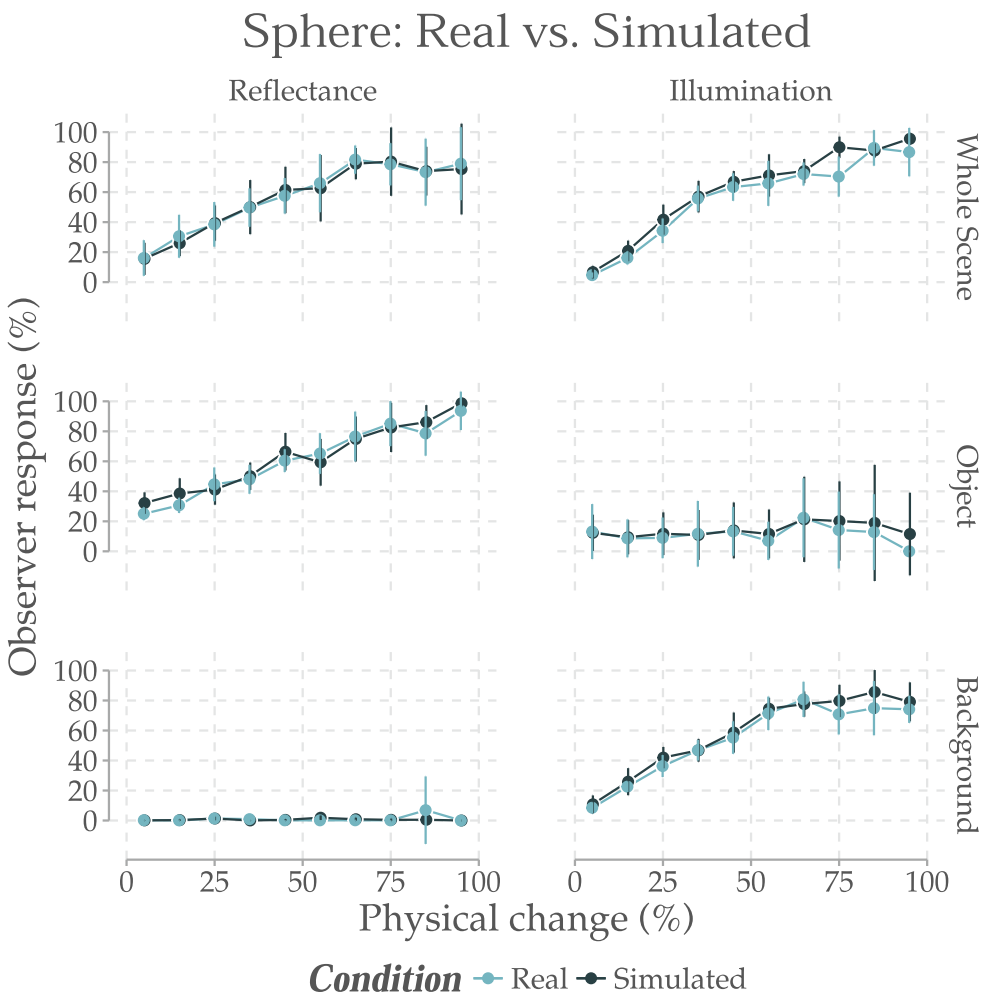

Fig. 10. Sphere scene: Observer responses for the two response types (illumination/reflectance) and the three views (whole scene, object only, or background only) parceled according to whether the image was a "real" hyperspectral image (blue) or a simulated image produced by the Mitsuba rendering system (black). It can be seen that observers performed similarly for both types of images, indicating that the rendering system includes the information that observers need to perform the task. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

could provide additional information for observers during the task. However, Fig. 11 shows that observers were already doing well with the spheres and that the Glavens did not appreciably improve performance. If anything, there seems to be a tendency for observers to overestimate the magnitude of changes when viewing the Glavens. However, there were no significant differences between GAMs fit separately to the data from the Glaven and the sphere conditions for each combination of response type (Illumination vs. Reflectance) and scene type (Whole Scene, Object, Background), as shown in Table 4. For the one case where $p = 0.05$ (top left panel of Fig. 11: perceived magnitude of illumination changes for the whole scene viewing condition), the BF still favors the simpler model of no strong difference between the two curves. Regardless, part of the apparent differences between the two conditions are likely due to noise and the few larger differences could be due to three alternative sources: (1) different observers did the spheres and Glavens experiments, (2) we reduced the overall number of trials for the Glavens experiments to save some time, and (3) the sphere data averages over the real and simulated scenes, whereas the Glavens were only tested with simulated scenes, so much more data contributes to the points for the sphere data.

4. Discussion

We deal with changes of illumination more frequently than we deal with changes of surface color, but both do occur in the natural world. Here, we have shown that observers are capable of perceptually

Table 3
Results from an ANOVA analysis comparing GAMs fit to the data shown in the six panels of Fig. 10. The models were nested, with the first fitting one smooth function to physical magnitude of stimulus change for both color axes and the second fitting a separate smooth function for each image condition ("real" vs. simulated). No significant differences were found between the two models at a $p < .05$ level, indicating, in general, no differences based on whether the image was simulated or not. The BFs consistently prefer the simpler model that does not fit a separate smooth function for each image condition.

	Illumination	Surface
Whole Scene	$F(86.35, 2.84) = NA$, $p = NA$, $BF = 188.51$	$F(91.80, 1.69) = NA$, $p = NA$, $BF = 34.45$
Object	$F(95.00, 1.00) = 0.516$, $p = .474$, $BF = 7.61$	$F(94.82, 1.18) = 0.232$, $p = .670$, $BF = 10.63$
Background	$F(86.40, 2.44) = 0.177$, $p = .877$, $BF = 66.24$	$F(91.64, 1.36) = 1.256$, $p = .279$, $BF = 6.28$

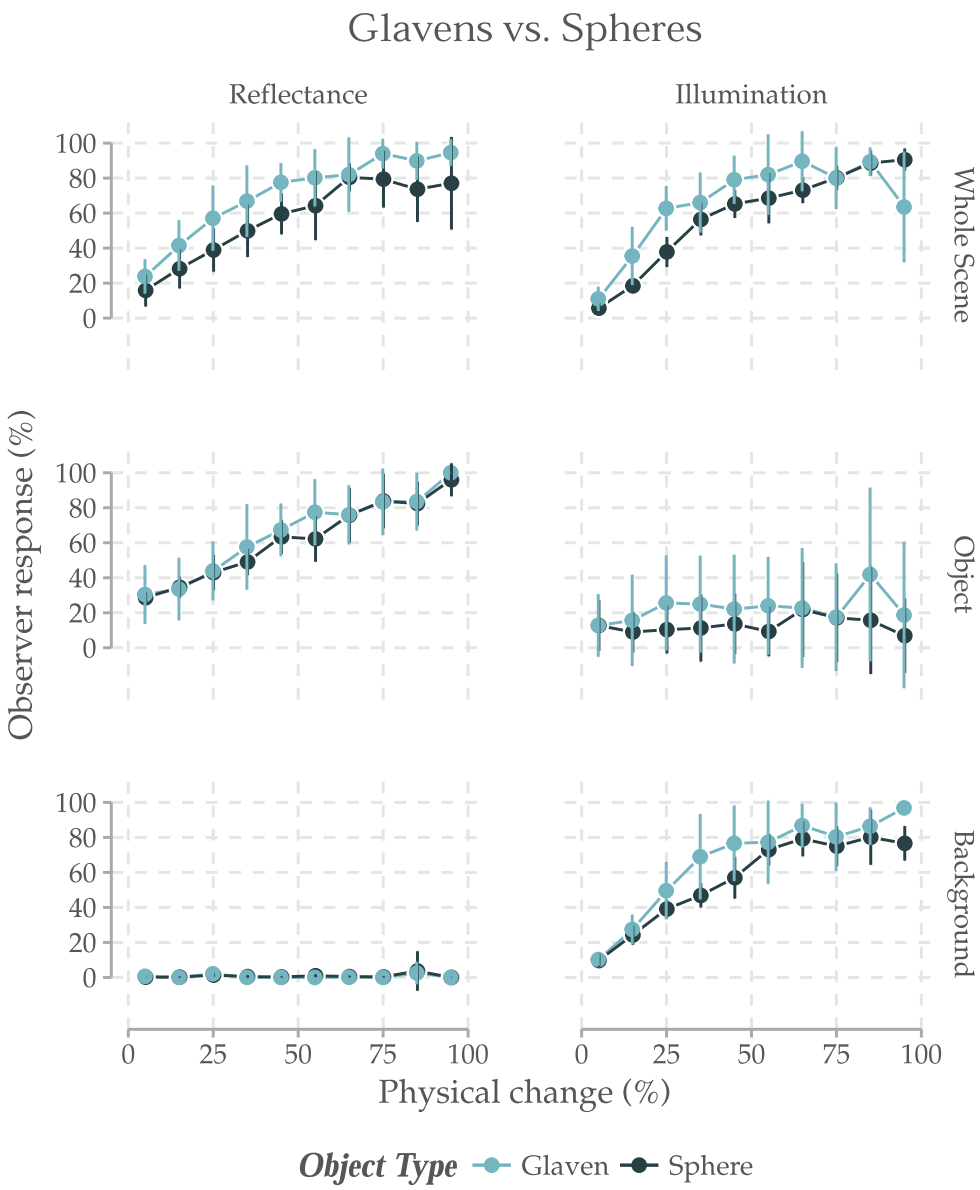

Fig. 11. Observer responses for the two response types (illumination/reflectance) and the three views (whole scene, object only, or background only) parceled according to whether the test object was a Glaven (blue) or a sphere (black). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Results from an ANOVA analysis comparing GAMs fit to the data shown in the six panels of Fig. 11. The models were nested, with the first fitting one smooth function to physical magnitude of stimulus change for both color axes and the second fitting a separate smooth function for each object type (Glaven vs. sphere). No significant differences were found between the two models at a $p < .05$ level, indicating, in general, no differences based on whether the object was a Glaven or a sphere. In the case of responses to illumination changes in the whole scene viewing condition, $p = .05$, which suggests a trend for Glavens and Spheres to be treated differently in that case, but the BF still favors the simpler model. For all other conditions, the BFs also consistently prefer the simpler model which does not fit a separate smooth function for each object type.

	Illumination	Surface
Whole Scene	$F(86.50, 2.87) = 2.750$, $p = .05$, BF = 3.13	$F(88.03, 2.14) = 0.223$, $p = .815$, BF = 38.41
Object	$F(92.00, 1.00) = 0.206$, $p = .651$, BF = 8.80	$F(91.00, 1.00) = 0.072$, $p = .788$, BF = 9.39
Background	$F(86.98, 2.77) = 1.146$, $p = .333$, BF = 29.09	$F(92.00, 1.00) = 0.696$, $p = .406$, BF = 6.82

disentangling simultaneous changes in surface and illumination. In particular, our task has been a new test of the robustness of various chromatic scene statistics commonly found in the color constancy literature. Localized versions of the statistics generally performed better than global versions of the statistics. One could argue that perhaps with a more complex scene, the current statistics would perform better, but we suggest that this is not the approach to take if one wants to better understand human behavior in our task. Regardless, we take our results as suggesting that observers are parsing scenes into a layered representation and performing color constancy computations on a local basis. While local color constancy computations are not a novel idea, the notion of global scene statistics seems to be more pervasive in the

field, if only implicitly. In fact, many of the mechanisms thought to be involved in color constancy are computed in a local manner at the level of neural implementations, which can assist in dealing with sudden changes in the pattern of the illumination when walking around (Werner, 2014).

In particular, our results contribute to the notion that the brain uses more than image-based measures to achieve color constancy. It is potentially parsing the scene into different layers, such as object, background, shadow, etc., and then using information in these layers. However, this is not a new idea (Anderson & Kim, 2009; Beck, 1972; Gilchrist et al., 1984). New avenues of investigation should continue to consider the information contained in different semantic regions of the scene (background, foreground, and shadows).

Another question that our results probe is whether or not observers need complicated shapes and articulated backgrounds to achieve color constancy in more natural settings. A matte sphere in a white box is about as simple a 3-D stimulus as one can get, being almost the 3-D analog of the classical center-surround stimulus, and yet, observers can handle the task. Giving them a more complex object, such as the Glavens, which have interreflections on their surface, did not produce marked changes in their behavior. To be clear, this is not to discredit previous studies which found that observers need more varied scenes to perform better at color constancy; it is already known that performance in a color constancy experiment is dependent on stimulus, task, and instructions (Foster, 2011; Hurlbert, 2007; Smithson, 2005; Werner, 2014). In fact, our observers are not at 100% performance in our task. We merely want to highlight that they are doing fairly well under rather reduced conditions. It will be interesting to investigate whether or not observers find the task more or less difficult with more complex scenes and objects.

Also of interest is the failure of the luminance-redness correlation and the relative success of the mean cone excitation ratios. The luminance-redness correlation has been previously proposed as a viable source of information about the illumination. It is found in some natural scenes and observers seem to use it to perform some tasks (Golz, 2008; Golz & MacLeod, 2002). However, it has been unable to fully explain observer judgments of the illumination change in our task. It is possible that our scenes may have been too simple for this statistic, but they also violated one of its principal expectations: that as the illuminant becomes brighter, it becomes redder, and vice versa. In our case, the illuminants were all fixed at roughly the same luminance and only changed in hue and saturation. As such, any change in the illuminant color towards or away from red was not accompanied by a concomitant change in luminance. In addition, a luminance-redness statistic will not be of assistance for color changes along an orthogonal 'blue-yellow' axis. Observers were using alternate information to perform the task and the luminance-redness correlation does not consider this.

The mean cone excitation ratios are also an interesting story. It is a statistic that was studied for the case where the illumination and the surface can change in color. As a result, it does better with accounting for observer behavior for both surface and illuminant changes than most statistics, but not perfectly (although, to be fair, no statistic was proposed as a perfect solution). In the work of Foster, Nascimento, Amano, and colleagues (Craven & Foster, 1992; Foster & Nascimento, 1994; Foster et al., 2000; Foster et al., 2006; Foster et al., 2001; Linnell & Foster, 1996), it has been consistently found that cone excitation ratios could serve as a physical invariant for color constancy, especially in tasks where the surface and illumination change simultaneously. However, the majority of the experiments with simultaneous changes have been done with 2-dimensional stimuli, while ours was 3-dimensional. While the work in Foster et al. (2006) used a 3-dimensional stimulus, observers were not asked to estimate the magnitude of change in either the illuminant or the test surface. Regardless, it seems that once these properties change simultaneously in a 3-dimensional scene and an observer must estimate their magnitude, cone excitation ratios can explain part of the variance, but not all. However, this is not to the detriment of cone excitation ratios. In Foster et al. (2006), it was found that cone excitation ratios explained 43.2% ± 14.5% of the variance on average and combinations with other statistics were necessary to account for more of the data, so it may only be part of the story.

This brings us to a theme that we have not dealt with in this paper: cue combination (Boyaci et al., 2006; Kraft, Maloney, & Brainard, 2002; Maloney, 2002; Yang & Maloney, 2001; Yang & Shevell, 2002). Cue combination suggests that observers use various sources of information to complete a task, giving these sources different weights as necessary to complete the task in an optimal fashion. One can represent this in a linear model framework and incorporate Bayesian priors for additional explanatory power. While work on color constancy has shown that observers are sensitive to manipulations of different information sources, it is still not clear when and how observers use each of these information sources. While our experiment does not probe this (further investigation is necessary), it is necessary to talk about the relevance of cue combination, since we only tested each statistic in isolation. We did this for two reasons: (1) many of the studies on these statistics considered them in isolation, so we wanted a fair comparison, and (2) finding the right combination of statistics to explain more than just the task at hand is a tricky matter and requires more conditions than we have tested in this study. It is also not clear if one ends up overfitting with respect to the stimulus. Aside from this, some of the chromatic scene statistics correlate with each other, and simply putting them into a linear regression and finding the best combination is not straightforward. In other forms of color constancy tasks, this is when one does want to use more complicated scenes because they make it possible to circumvent these correlations to a certain degree, but this does not fully solve the problem for our paradigm. For example, a scene with white walls and a black floor could reduce the correlation between the White patch and Average color statistics for a single image, because one can change the contrast between the walls and the floor, which would change the result of the White patch statistic, but leave the Average color statistic unchanged. However, our task would probably become too difficult if the walls also changed color. In addition, our task is about estimating magnitude of color change, and even if the estimates of the White patch and Average color statistics were de-correlated for each pair of images on a trial, their estimates of magnitude change would still highly correlate. Please note that we wish to make it clear that we do think that observers combine cues, but the question of how observers make use of different statistics in different color constancy tasks is still open.

It is also worthwhile to consider that a statistic which is good for estimating the illuminant may not be what one wants to use to estimate surface color changes and vice versa. For example, the Average color statistic is known to not be a viable statistic to depend on for illuminant estimation, as mentioned in the Results section. However, it does do well at predicting observer behaviour for perceived magnitude of surface color changes in our task. It is possible that an observer might use a different statistic to estimate the illumination in most circumstances, but then use the Average color statistic to estimate surface color. This would parallel previous work showing that observers seem to use the mean color to categorize images of leaves into 'red' and 'green' color categories (Milojevic, Ennis, Toscani, & Gegenfurtner, 2018), although categorization and estimation of color change magnitudes are two different kinds of tasks which might use different types of information. On the other hand, the White patch statistic has been linked to observer behaviour in other tasks (Giesel & Gegenfurtner, 2010; Gilchrist et al., 1999; Toscani et al., 2013) and is also doing well at predicting observers' illumination magnitude change responses in our task. It could be that observers use the White patch statistic to estimate the illumination, then use that to do a color constancy correction, and then use the Average color statistic to estimate the surface color. This strategy would also be useful in the presence of specular highlights on glossy objects, since highlights are typically composed of the illuminant color and are typically very bright, while the object's more diffuse reflections

Table 5
Table showing the average number of bins where the standard deviation of observer responses does not change appreciably between the whole scene and the object only viewing conditions, as well as the average change in standard deviation between the two conditions for the bins that did not satisfy the criteria. The criterion was defined as: bins where standard deviation for object only viewing condition was less than the whole scene viewing condition and/or bins with a difference in standard deviation of 5% or less between the two conditions. As can be seen, usually about half of the bins do not change appreciably in variability, but those that do, change substantially, indicating that observers are having a tougher time with the object only condition and are probably just perceiving any illumination change as part of the reflectance change, causing them to under- or over-estimate the actual reflectance change.

	Avg. bins w/unchanged std. dev.	Avg. change in std. dev. for remaining bins
Spheres - Hyperspectral images	6.2 ± 1.33	12.75% ± 5.47%
Spheres - Rendered images	5.36 ± 1.35	11.72% ± 5.03%
Glavens	5.01 ± 2.04	12.36% ± 7.95%

would be diagnostic of the body color and are darker than the highlights (Shafer, 1985). However, more work should be done to test these concepts more fully, including testing objects with more complex gradations of color across their surfaces, especially with respect to estimations of the magnitude of surface color changes. It will also be of interest to determine how to incorporate cone excitation ratios as part of the process.

Lastly, we wish to bring our readers' attention to the object-only viewing condition for both the spheres and the Glavens. For this condition, observers were essentially correct on average about the change in surface reflectance, but not about the change in the illuminant. To us, this seems counter-intuitive, since the classical idea of color constancy is that one first determines the color of the illuminant, then "subtracts" this from the scene, thereby "discounting" it (Helmholtz, 1867) and obtaining stable estimates of surface properties. Reasoning from this condition only, it would seem then that "discounting" the illuminant really does mean to completely discard all of the information about the illuminant and forget about it, but this does not hold, since we see illuminant colors in our daily lives and observers were able to extract the illuminant color in both the whole scene and background only viewing conditions. One could alternatively explain the data by suggesting that when observers see the object only condition, they will perceive any color change as a reflectance only change, so any contribution of the illuminant change would contaminate their judgment and cause them to under- or over-estimate the actual reflectance change. If this were the case, then they should be roughly correct on average, as we find, but the variability of their responses should be larger in the object only condition. We find this basically to be the case; see Table 5, where the average change in the standard deviation between the two conditions is shown. At least for the objects that we tested here, it seems to be the case that an observer needs a background for accurate and reliable illuminant estimation.

5. Conclusions

In conclusion, we have found that observers can deal with simultaneous changes in surface reflectance and illumination. We propose that further investigation be done to tackle the following three items: (1) determine how the visual system uses information from a potential layer decomposition of the scene, (2) determine which statistic(s) are extracted from each of these layers, and (3) determine how they are synthesized to produce a final estimate of illumination/surface color change. While none of these are new objectives in the field of color constancy research, we merely wish to reiterate their simultaneous importance. Lastly, it will be interesting to see what happens when other materials are tested.

Acknowledgments

We would like to thank Anya Hurlbert for her suggestion of comparing the different color axes. We would like to thank the following for their comments and suggestions on earlier versions of the paper: Sylvia Pont, Anya Hurlbert, Qasim Zaidi, Arthur Shapiro, David Brainard, Karl Gegenfurtner, and Matteo Toscani. Funding was provided by the Alexander von Humboldt Foundation in the framework of the Sof'ja Kovalevskaja Award endowed by the German Federal Ministry of Education and Research.

References

Amano, K., & Foster, D. (2004). Colour constancy under simultaneous changes in surface position and illuminant. *Proceedings of the Royal Society of London Series B, 271*(1555), 2319–2326. https://doi.org/10.1098/rspb.2004.2884.
Anderson, B. L., & Kim, J. (2009). Image statistics do not explain the perception of gloss and lightness. *Journal of Vision, 9*(11), 1–17. https://doi.org/10.1167/9.11.10.
Beck, J. (1972). *Surface color perception.* Ithaca, New York: Cornell University Press.
Bloj, M. G., Kersten, D., & Hurlbert, A. C. (1999). Perception of three-dimensional shape influences colour perception through mutual illumination. *Nature, 402*(6764), 877–879. https://doi.org/10.1038/47245.
Boyaci, H., Doerschner, K., & Maloney, L. T. (2006). Cues to an equivalent lighting model. *Journal of Vision, 6*, 106–118. https://doi.org/10.1167/6.2.2.
Boyaci, H., Doerschner, K., Snyder, J. L., & Maloney, L. T. (2006). Surface color perception in three-dimensional scenes. *Visual Neuroscience, 23*(3–4), 311–321. https://doi.org/10.1017/S0952523806233431.
Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.
Brainard, D. H., Kraft, J. M., & Longere, P. (2003). Color constancy: Developing empirical tests of computational models. *Colour perception: Mind and the physical world* (pp. 307–334). Oxford University Press.
Brown, R. O., & MacLeod, D. I. (1997). Color appearance depends on the variance of surround colors. *Current Biology, 7*(11), 844–849.
Buchsbaum, G. (1980). A spatial processor model for object colour perception. *Journal of the Franklin Institute, 310*(1), 1–26. https://doi.org/10.1016/0016-0032(80)90058-7.
Burnham, R. W., Evans, R. M., & Newhall, S. M. (1957). Prediction of color appearance with different adaptation illuminations. *Journal of the Optical Society of America, 47*, 35–42. https://doi.org/10.1364/JOSA.47.000035.
Craven, B. J., & Foster, D. H. (1992). An operational approach to colour constancy. *Vision Research, 32*(7), 1359–1366. https://doi.org/10.1016/0042-6989(92)90228-B.
Delahunt, P. B., & Brainard, D. H. (2004). Does human color constancy incorporate the statistical regularity of natural daylight? *Journal of Vision, 4*, 57–81. https://doi.org/10.1167/4.2.1.
Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of Physiology, 357*, 241–265. https://doi.org/10.1111/(ISSN)1469-7793.
Ennis, R., Schiller, F., Toscani, M., & Gegenfurtner, K. R. (2018). Hyperspectral database of fruits and vegetables. *Journal of the Optical Society of America, 35*(4), B256–B266. https://doi.org/10.1364/JOSAA.35.00B256.
Fairchild, M. D., & Reniff, L. (1995). Time course of chromatic adaptation for color-appearance judgements. *Journal of the Optical Society of America A, 12*, 824–833. https://doi.org/10.1364/JOSAA.12.000824.
Foster, D. H. (2011). Color constancy. *Vision Research, 51*(7), 674–700. https://doi.org/10.1016/j.visres.2010.09.006.
Foster, D. H., Amano, K., & Nascimento, S. M. C. (2000). How temporal cues can aid colour constancy. *Color Research and Application, 26*, S180–S185.
Foster, D. H., Amano, K., & Nascimento, S. M. C. (2006). Color constancy in natural scenes explained by global image statistics. *Vision Research, 23*, 1–17. https://doi.org/10.1017/S0952523806233455.
Foster, D. H., Amano, K., & Nascimento, S. M. C. (2016). Time-lapse ratios of cone excitations in natural scenes. *Vision Research, 120*, 45–60. https://doi.org/10.1016/j.visres.2015.03.012.
Foster, D. H., & Nascimento, S. M. C. (1994). Relational color constancy from invariant cone-excitation ratios. *Proceedings of the Royal Society of London Series B-Biological Sciences, 257*(1349), 115–121.
Foster, D. H., Nascimento, S. M., Amano, K., Arend, L., Linnell, K. J., Nieves, J. L., ... Foster, J. S. (2001). Parallel detection of violations of color constancy. *Proceedings of the National Academy of Sciences of the United States of America, 98*, 8151–8156. https://doi.org/10.1073/pnas.141505198.
Funt, B., & Drew, M. (1993). Color space analysis of mutual illumination. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 15(12), 1319–1326. https://doi.org/10.1109/34.250838.
Funt, B. V., Drew, M. S., & Ho, J. (1991). Color constancy from mutual reflection. *International Journal of Computer Vision, 6*, 5–24. https://doi.org/10.1007/BF00127123.
Gerhard, H. E., & Maloney, L. T. (2010). Detection of light transformations and concomitant changes in surface albedo. *Journal of Vision, 10*(9), 1–14. https://doi.org/10.1167/10.9.1.
Giesel, M., & Gegenfurtner, K. R. (2010). Color appearance of real objects varying in material, hue, and shape. *Journal of Vision, 10*(9), 1–21. https://doi.org/10.1167/10.9.10.
Gilchrist, A., & Jacobsen, A. (1984). Perception of lightness and illumination in a world of one reflectance. *Perception, 13*(1), 5–19. https://doi.org/10.1068/p130005.
Gilchrist, A., Kossyfidis, C., Bonato, F., Agostini, T., Cataliotti, J., Li, X., ... Economou, E. (1999). An anchoring theory of lightness perception. *Psychological Review, 106*, 795–834.
Golz, J. (2008). The role of chromatic scene statistics in color constancy: Spatial integration. *Journal of Vision, 8*(13), 1–16. https://doi.org/10.1167/8.13.6.
Golz, J., & MacLeod, D. I. A. (2002). Influence of scene statistics on colour constancy. *Nature, 415*(6872), 637–640. https://doi.org/10.1038/415637a.
Hansen, T., & Gegenfurtner, K. R. (2013). Higher order color mechanisms: Evidence from noise-masking experiments in cone contrast space. *Journal of Vision, 13*(1), 1–21. https://doi.org/10.1167/13.1.26.
Helmholtz, H. (1867). *Handbuch der physiologischen Optik.* Leipzig: Voss.
Hering, E. (1878). *Grundzüge der Lehre vom Lichtsinn.* Berlin: Springer.
Hurlbert, A. (1998). *Computational models of color constancy.* Cambridge University Press283–321.
Hurlbert, A. (1999). Colour vision: Is colour constancy real? *Current Biology, 9*, R558–R561. https://doi.org/10.1016/S0960-9822(99)80354-6.
Hurlbert, A. (2007). Colour constancy. *Current Biology, 17*(21), R906–R907. https://doi.org/10.1016/j.cub.2007.08.022.
Jakob, W. (2017). Mitsuba renderer. https://www.mitsuba-renderer.org/.
Jameson, D., & Hurvich, L. M. (1989). Essay concerning color constancy. *Annual Review of Psychology, 40*, 1–22.
Kaiser, P. K., & Boynton, R. M. (1996). *Human color vision.* Washington, DC: Optical Society of America.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572.
Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? In *Perception 36 ECVP abstract supplement* (Vol. 36, pp. 1–16).
Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in R.* New York: Springerhttps://doi.org/10.1007/978-1-4614-4475-6.
Koenderink, J. J., & van Doorn, A. J. (1983). Geometrical modes as a general method to treat diffuse interreflections in radiometry. *Journal of the Optical Society of America, 73*(6), 843–850.
Kraft, J. M., & Brainard, D. H. (1999). Mechanisms of color constancy under nearly natural viewing. *Proceedings of the National Academy of Sciences, 96*(1), 307–312. https://doi.org/10.1073/pnas.96.1.307.
Kraft, J. M., Maloney, S. I., & Brainard, D. H. (2002). Surface-illuminant ambiguity and color constancy: Effects of scene complexity and depth cues. *Perception, 31*(2), 247–263.
Land, E. H., & McCann, J. J. (1971). Lightness and retinex theory. *Journal of the Optical Society of America, 61*(1), 1–11. https://doi.org/10.1364/JOSA.61.000001.
Langer, M. S. (1999). When shadows become interreflections. *International Journal of Computer Vision, 34*(2/3), 193–204.
Langer, M. S. (2001). A model of how interreflections can affect color appearance. In *Proceedings of the International Colour Vision Society* (Vol. 26, pp. S218–S221).
Linnell, K. J., & Foster, D. H. (1996). Dependence of relational colour constancy on the extraction of a transient signal. *Perception, 25*, 221–228. https://doi.org/10.1068/p250221.
MacLeod, D. I., & Boynton, R. M. (1979). Chromaticity diagram showing cone excitation by stimuli of equal luminance. *Journal of the Optical Society of America, 69*(8), 1183–1186.
Maloney, L. T. (1999). *Physics-based approaches to modeling surface color perception.* Cambridge, UK: Cambridge University Press387–422.
Maloney, L. T. (2002). Illuminant estimation as cue combination. *Journal of Vision, 2*, 493–504. https://doi.org/10.1167/2.6.6.
Matsakis, N. D., & Klock, F. S., II (2014). *The rust language, Vol. 34*https://doi.org/10.1145/2692956.2663188.
Milojevic, Z., Ennis, R., Toscani, M., & Gegenfurtner, K. R. (2018). Categorizing natural color distributions. *Vision Research*. https://doi.org/10.1016/j.visres.2018.01.008.
Moon, P. (1940). On interreflections. *Journal of the Optical Society of America, 30*(5), 195–205. https://doi.org/10.1364/JOSA.30.000195.
Nascimento, S. M., & Foster, D. H. (2000). Relational color constancy in achromatic and isoluminant images. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 17*, 225–231.
Nascimento, S. M. C., & Foster, D. H. (2001). Detecting changes of spatial cone-excitation ratios in dichoptic viewing. *Vision Research, 41*(20), 2601–2606. https://doi.org/10.1016/S0042-6989(01)00142-0.
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), 1–10. https://doi.org/10.1371/journal.pone.0087989.
Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.
Phillips, F., Egan, E. J. L., & Perry, B. N. (2009). Perceptual equivalence between vision and touch is complexity dependent. *Acta Psychologica, 132*(3), 259–266. https://doi.org/10.1016/j.actpsy.2009.07.010.
Phillips, F., Casella, M. W., & Egan, E. J. L. (2016). Glaven objects (v1.4) [3D Object Files]. https://academics.skidmore.edu/blogs/flip/?page_id=813.
Rinner, O., & Gegenfurtner, K. (2000). Time course of chromatic adaptation for color appearance and discrimination. *Vision Research, 40*, 1813–1826. https://doi.org/10.1016/S0042-6989(00)00050-X.
Ruppertsberg, A. I., & Bloj, M. (2007). Reflecting on a room of one reflectance. *Journal of Vision, 7*(13), 1–13. https://doi.org/10.1167/7.13.12.
Schanda, J. (2014). *CIE chromaticity diagrams, CIE purity, CIE dominant wavelength.* Heidelberg, Berlin, Heidelberg: Springer, Berlin1–6. https://doi.org/10.1007/978-3-642-27851-8_325-1.
Shafer, S. A. (1985). Using color to separate reflection components. *Color Research and Application, 10*(4), 210–218. https://doi.org/10.1002/col.5080100409.
Smithson, H. E. (2005). Sensory, computational and cognitive components of human colour constancy. *Philosophical Transactions of the Royal Society B, 360*(1458), 1329–1346. https://doi.org/10.1098/rstb.2005.1633.
Stockman, A., & Sharpe, L. T. (2000). The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision Research, 40*(13), 1711–1737.
Stockman, A., Sharpe, L. T., & Fach, C. (1999). The spectral sensitivity of the human short-wavelength sensitive cones derived from thresholds and color matches. *Vision Research, 39*, 2901–2927.
The Rust Programming Language (2017). https://www.rust-lang.org.
Toscani, M., Valsecchi, M., & Gegenfurtner, K. R. (2013). Optimal sampling of visual information for lightness judgments. *Proceedings of the National Academy of Sciences, 110*(27), 11163–11168. https://doi.org/10.1073/pnas.1216954110.
Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779–804. https://doi.org/10.3758/BF03194105.
Ward, G. (1989). The RADIANCE lighting simulation and rendering system. In *21st annual conference on computer graphics and interactive techniques* (pp. 459–472). arXiv:1011.1669v3. https://doi.org/10.1145/192161.192286.
Webster, M. A., & Mollon, J. D. (1995). Colour constancy influenced by contrast adaptation. *Nature, 373*, 694–698. https://doi.org/10.1038/373694a0.
Werner, A. (2014). Spatial and temporal aspects of chromatic adaptation and their functional significance for colour constancy. *Vision Research, 104*, 80–89. https://doi.org/10.1016/j.visres.2014.10.005.
Yang, J. N., & Maloney, L. T. (2001). Illuminant cues in surface color perception: Tests of three candidate cues. *Vision Research, 41*(20), 2581–2600. https://doi.org/10.1016/S0042-6989(01)00143-2.
Yang, J. N., & Shevell, S. K. (2002). Stereo disparity improves color constancy. *Vision Research, 42*(16), 1979–1989.
Zaidi, Q., & Halevy, D. (1993). Visual mechanisms that signal the direction of color changes. *Vision Research, 33*(8), 1037–1051.