Behavior Research Methods
https://doi.org/10.3758/s13428-024-02352-0

NOTES/COMMENTS/REPLY MANUSCRIPT

Lightness constancy in reality, in virtual reality, and on flat-panel displays

Khushbu Y. Patel[1] · Laurie M. Wilcox[1] · Laurence T. Maloney[2] · Krista A. Ehinger[3] · Jaykishan Y. Patel[1] · Emma Wiedenmann[1,4] · Richard F. Murray[1]

Accepted: 30 January 2024

Abstract
Virtual reality (VR) displays are being used in an increasingly wide range of applications. However, previous work shows that viewers often perceive scene properties very differently in real and virtual environments and so realistic perception of virtual stimuli should always be a carefully tested conclusion, not an assumption. One important property for realistic scene perception is surface color. To evaluate how well virtual platforms support realistic perception of achromatic surface color, we assessed lightness constancy in a physical apparatus with real lights and surfaces, in a commercial VR headset, and on a traditional flat-panel display. We found that lightness constancy was good in all three environments, though significantly better in the real environment than on the flat-panel display. We also found that variability across observers was significantly greater in VR and on the flat-panel display than in the physical environment. We conclude that these discrepancies should be taken into account in applications where realistic perception is critical but also that in many cases VR can be used as a flexible alternative to flat-panel displays and a reasonable proxy for real environments.

Keywords Vision · Lightness · Constancy · Virtual reality · Realism

Introduction

In recent years, there have been rapid advances in virtual reality (VR) technology (Greengard, 2019; Zhan et al., 2020; Xiong et al., 2021), and VR has found a wide range of recreational and professional uses, including gaming (Linowes, 2020), cinema (Marantz, 2016), training (Xie et al., 2021), physical rehabilitation (Elor & Kurniawan, 2020), psychological therapy (Ong et al., 2022), and vision research (Hibbard, 2023; Scarfe & Glennerster, 2015, 2019). In many applications, key goals for VR design include *immersion*, the ability of the VR environment to provide rich, interactive stimuli, and *presence*, the viewer's subjective sense of being fluidly engaged in the environment (Berkman & Akan, 2019; Sanchez-Vives & Slater, 2005; Slater et al., 2009). In some applications, an additional goal is *realism*, meaning that people and things should appear to the viewer in VR just as they would appear in the real world (Jung & Lindeman, 2021). Realism is important, for example, in some kinds of training, where the participant learns skills that must transfer to the real world, and it is also important in vision research, where VR is often used to show precisely controlled, computer-generated stimuli that are stand-ins for real objects and scenes. Interestingly, the available evidence suggests that these three goals are only loosely related, e.g., a high degree of realism may not be a prerequisite for strong presence (Jung & Lindeman, 2021; Slater, 2018).

Here we evaluate a specific but fundamental aspect of realism in VR, namely the extent to which viewers perceive black, white, and grey surface colors in virtual environments with the same accuracy as in real environments. We report achromatic color matching experiments with a physical apparatus that uses real lights and surfaces, a VR environment designed to simulate the physical apparatus, and a flat-panel monitor that shows a rendered image of the physical apparatus. Our goal is to quantify and compare perception of achromatic surface color in these three environments, in order

✉ Khushbu Y. Patel
khushbuypatel1234@gmail.com

1 Department of Psychology and Centre for Vision Research, York University, Toronto, Canada

2 Department of Psychology, New York University, New York, USA

3 School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

4 Department of Psychology, Carl Von Ossietzky Universität Oldenburg, Oldenburg, Germany

to evaluate the extent to which VR and flat-panel displays can substitute for real lights and surfaces in applications where realism is important.

Lightness perception

Visual processing begins with stimulation of the 2D array of photoreceptors in the retina, but the retinal image provides only indirect and ambiguous information about useful properties of things we see (Murray, 2021; Belhumeur et al., 1999). For example, it is often useful to know the *surface color* of things in the environment: this apple is green, that fox is light grey, and so on. The retinal image, though, depends both on the intrinsic surface colors of objects (i.e., the proportion of light they reflect at various wavelengths), and also on the spectrum and intensity of illumination in the environment. In order to perceive surface color accurately, the human visual system must disentangle the effects of surface color and lighting. This is a difficult computational problem that is still not well understood (Barron & Malik, 2015; Li et al., 2020), but remarkably, the human visual system usually accomplishes it with a high degree of accuracy (Murray, 2021; Brainard & Maloney, 2011). This ability to perceive surface color accurately in a wide range of lighting conditions and environments is called *color constancy*.

In work on surface color perception and color constancy, researchers sometimes simplify the problem under study by limiting stimuli to achromatic colors: black, white, and shades of grey. In this case, the surface property of interest is *reflectance*, the proportion of incident light in the visible wavelength region that is reflected by a surface[1]. In principle, reflectance can range from 0.0 to 1.0, but in everyday materials, black surfaces have a reflectance around 0.03, white surfaces have a reflectance around 0.9, and grey surfaces have values in between.

Reflectance is a physical property. The corresponding perceptual property, *lightness*, is defined[2] as perceived reflectance (Gilchrist, 2006). For example, in a printed copy of the snake illusion (Fig. 1), the four diamonds have the same physical reflectance, but very different lightness, as the top two diamonds appear much lighter than the bottom two. Lightness perception is a rich domain with a long history in experimental psychology; for reviews, see Gilchrist (2006) and Murray (2021).

[1] A more precise definition of reflectance takes into account the spectral distribution of the illumination and the spectral sensitivity of the human visual system. We will skirt details of photometry and colorimetry when they are not central to the discussion. For more information, see McCluney (1994) and Fairchild (2013).

[2] Alternatively, lightness is sometimes defined in terms of the perceived intensity of light reflected by a surface, relative to the light reflected by a white surface (Fairchild, 2013 p.88). We will follow the definition in the main text.

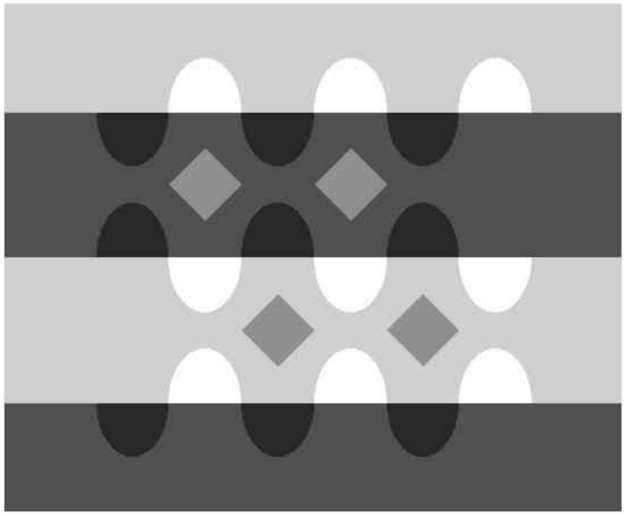

Fig. 1 The snake illusion. The four diamonds are physically identical, but the top two appear lighter than the bottom two. Adapted with permission from Adelson (2000), copyright The MIT Press

Lightness perception raises many of the same problems as color perception more generally. When we view a grey surface, the retinal light intensity depends on both the surface's reflectance and the amount of light incident on the surface. For a matte (i.e., Lambertian) surface, the luminance (l) seen by an observer is proportional to the surface's reflectance (r) and to the illuminance (i) incident on the surface.

$$l = \frac{ir}{\pi} \tag{1}$$

(The factor of π follows from the definitions of SI units for luminance and illuminance.) Clearly, many combinations of illuminance i and reflectance r can produce any given luminance l. Nevertheless, the human visual system is able to use contextual information and prior knowledge to estimate surface reflectance from such ambiguous luminance measurements. This ability is called *lightness constancy*.

Asymmetric lightness matching

A common experimental method for studying lightness perception is asymmetric lightness matching. In this design, the observer is shown a grey reference patch under one illuminant, and an adjustable match patch under another illuminant[3] (Fig. 2(a)). The observer's task is to adjust the reflectance of the match patch so that it appears to be the same as the reflectance of the reference patch. An observer who has perfect lightness constancy will make this match setting accurately, even though the two patches are shown under different

[3] This is 'asymmetric' matching because the reference and match stimuli are viewed under different illuminants.

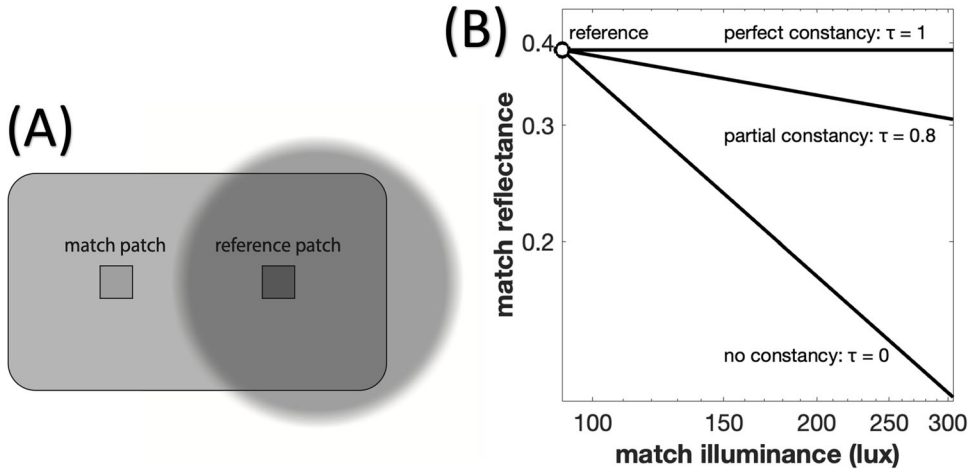

Fig. 2 A typical asymmetric lightness matching experiment. **a** The reference patch has a fixed reflectance and illumination level. The match patch has an adjustable reflectance, and is shown under a different illumination level. The observer adjusts the match patch so that it appears to have the same reflectance as the reference patch. **b** Lightness matches for three different Thouless ratios. At high Thouless ratios (good constancy), the match reflectance is not strongly affected by the illuminance at the match patch. At low Thouless ratios (poor constancy), the match reflectance decreases substantially as illuminance increases

illuminants. One who has poor lightness constancy, and for example matches the luminance of the two patches instead of the reflectance, will make inaccurate match settings. Thus the observer's match settings can be used to evaluate their lightness constancy under various experimental conditions.

One metric for quantifying lightness constancy is the *Brunswik ratio* (Brunswik, 1928). Suppose that in an asymmetric lightness matching task, the correct reflectance match setting is r_1 (indicating perfect constancy), and the reflectance setting that would result from luminance matching is r_0 (indicating a complete lack of constancy). If the observer's actual match setting is r_m, then their Brunswik ratio is

$$\beta = \frac{r_m - r_0}{r_1 - r_0} \tag{2}$$

This ratio indicates where the match setting r_m lies between the two theoretically defined matches r_0 and r_1. An observer who matches luminance has $\beta = 0$, and one who has perfect lightness constancy has $\beta = 1$. Matches resulting from partial lightness constancy have $0 < \beta < 1$.

The *Thouless ratio* is a modification of the Brunswik ratio that takes into account the perceptual scaling of lightness (Thouless, 1931). Many studies have shown that lightness is a compressive function of reflectance, e.g., the perceived difference between reflectances 0.10 and 0.15 is much larger than the perceived difference between 0.60 and 0.65. In the Munsell color system and CIELAB color space, for example, which are attempts at perceptually uniform spaces where perceived color differences are proportional to color coordinate differences, the lightness dimension is approximately proportional to the cube root (a compressive function) of the corresponding physical dimension (Fairchild 2013 pp. 80, 100). Similarly, the Weber–Fechner law states that a perceived quantity is a logarithmic function of the corresponding physical quantity (Fechner, 1860/1966) – again, a compressive mapping. The Thouless ratio uses a logarithmic transform of reflectance and is defined as

$$\tau = \frac{\log r_m - \log r_0}{\log r_1 - \log r_0} \tag{3}$$

Like the Brunswik ratio, the Thouless ratio indicates where the observer's match setting lies between two theoretically defined reference points, corresponding to perfect constancy and no constancy. It is also similar to the 'color constancy index' that indicates where an observer's color match setting lies relative to two reference points in an approximately perceptually uniform color space (Arend et al., 1991; Brainard, 1998). The values of Thouless ratios found in lightness matching experiments depend strongly on stimulus properties such as the complexity of the scene, but typical values in rich, naturalistic scenes are around 0.8 (Patel et al., 2018).

In an asymmetric lightness matching task such as the one we report below, the observer views a reference patch with a fixed reflectance and illumination, and a match patch under illumination that varies from trial to trial. The observer adjusts the reflectance of the match patch until it appears to be the same as the reflectance of the reference patch. The match setting r_m can be plotted as a function of the illuminance i_m at the match patch (Fig. 2(b)). In Appendix A we show that for an observer with a fixed Thouless ratio τ, this plot is a straight line on log-log axes, with slope $\tau - 1$. Thus such a plot gives a straightforward visual representation

of the observer's degree of lightness constancy. Two useful benchmarks are the line for perfect constancy ($\tau = 1$), with slope zero, and the line for luminance matching ($\tau = 0$), with slope -1.

Previous work

Experiments on visual perception often use computer-generated stimuli in order to maintain precise control over key image features. In fact, some experimental designs would be difficult to implement at all without computer-generated images, such as those using inconsistent lighting cues (Wilder et al., 2019; Ostrovsky et al., 2005) or inconsistent depth cues (Landy et al., 1995), or using scenes rendered without inter-reflections (Bloj et al., 1999). VR extends this control further and allows the experimenter to create immersive and interactive but well-controlled virtual environments.

But if we simulate a visual property such as surface reflectance in a computer-generated scene, does it generate the same visual percept as in a real scene? Clearly the answer will depend on the details of rendering and display methods. Previous work suggests, though, that even when care is taken to render stimuli realistically, perception of real and virtual environments can differ substantially, and realistic perception of virtual stimuli should always be a carefully tested conclusion, not an assumption. Some previous work shows that observers use qualitatively similar strategies in real and virtual environments, but with important quantitative differences. For example, Kimura et al. (2017) examined how observers determine their own orientation within real and VR environments, and found that although they rely on geometric cues and familiar landmarks in both cases, they rely more strongly on landmarks in VR. Rzepka et al. (2023) also found that when judging size, observers integrate familiar size cues differently in reality and in VR. Hartle and Wilcox (2022) found that observers make qualitatively similar depth judgements in real and virtual scenes but fail to achieve depth constancy in VR due to conflict between vergence and accommodation cues. Creem-Regehr et al. (2022) review work showing that observers underestimate egocentric distance in virtual environments compared to real environments.

Work on lightness and color constancy has also revealed differences between perception of real and virtual environments. Bloj et al. (2004) used real objects and lights to examine whether observers compensate for local lighting conditions when estimating the reflectance of surface patches at various 3D orientations, and Boyaci et al. (2003) independently reported a similar study using computer-generated stimuli viewed through a stereoscope. These two studies reported qualitatively similar results and showed that observers take into account the 3D distribution of lighting when estimating reflectance of surface patches at various 3D orientations. However, Morgenstern et al. (2014) reanalyzed their results in a way that made them quantitatively comparable and showed that lightness constancy was much better in Bloj et al.'s task, which used real stimuli, than in Boyaci et al.'s, which used computer-generated stimuli. This comparison had the caveat that the scenes in these two independently designed studies were broadly similar but different in many details. Patel et al. (2018) also found substantial differences between lightness judgements in real and computer-generated scenes. On the other hand, Blakeslee et al. (2008) examined lightness judgements in real scenes and in virtual reality, using 2D Mondrian stimuli, and found similar results in both cases. Similarly, Gil Rodríguez et al. (2022) found good color constancy in a complex scene shown in a VR headset, though without measuring constancy in a corresponding real scene for direct comparison. Radonjić et al. (2016) found that illumination discrimination thresholds were practically identical in a stereoscope and in a physical apparatus.

None of these previous studies compared measures of lightness constancy in real scenes and in rich VR environments. Here we carried out this comparison. We measured performance on a lightness matching task in three separate conditions, using (a) real lights and surfaces, (b) a VR headset, and (c) a flat-panel monitor, with key stimulus properties carefully matched across conditions. The real or simulated apparatus was a simple 2D panel with a single shadow boundary down the middle. We chose this configuration in order to examine a relatively simple lightness matching task (e.g., no requirement to estimate 3D surface orientation), and to test whether lightness constancy can be as good in VR as in a real environment under favorable conditions. We examined achromatic color perception as a special case of color constancy, where observers must overcome the ambiguity of greyscale retinal images in order to judge surface reflectances of objects being viewed in moderately complex, realistic scenes.

Methods

Participants

There were 12 participants, all of whom were unaware of the purpose of the experiment. Ten were female, two were male, and ages ranged from 19 to 35 years. Participants gave written informed consent and were paid for their participation. All reported normal or corrected-to-normal acuity and reported

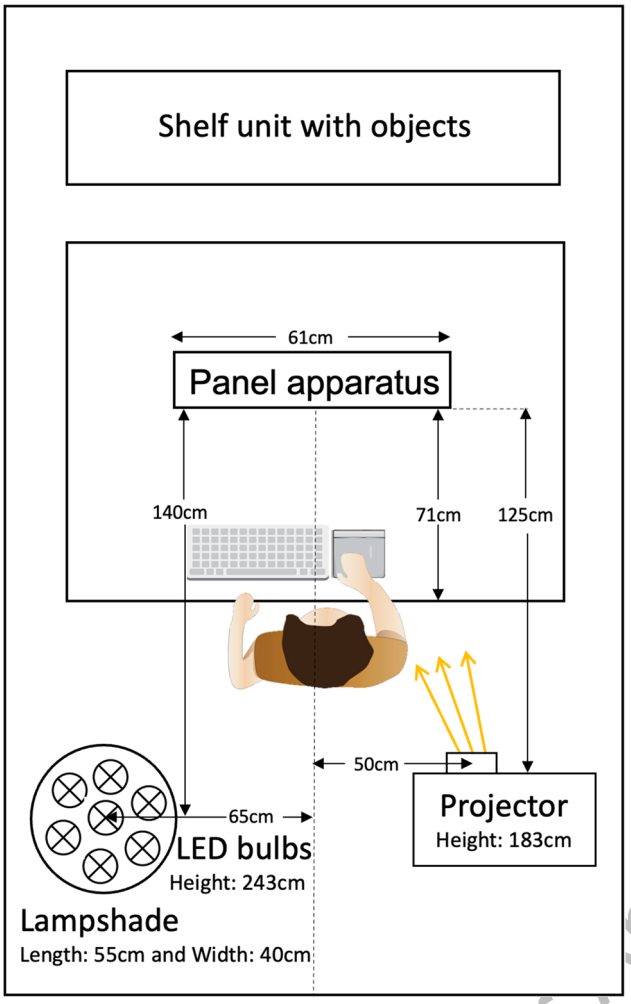

Fig. 3 A schematic top-down view of the testing room and apparatus in the physical environment

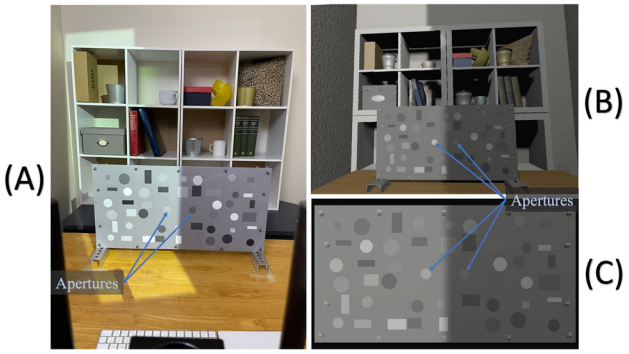

Fig. 4 **a** Photograph of the physical room and apparatus, taken from the viewing position used by observers in the experiment. **b** A screen capture of the VR environment. **c** The panel apparatus rendered and displayed on a flat-panel monitor

no known anomalies in color vision. All procedures were approved by the Office of Research Ethics at York University.

Stimuli

Physical environment The observer sat in a 3.00 m × 1.75 m room and viewed the experimental apparatus on a table (Fig. 3). The front of the apparatus was an achromatic paper panel, 61 cm horizontal × 31 cm vertical, at a viewing distance of 71 cm (Figs. 3 and 4). The panel showed a printed image with a background reflectance of 0.35 (i.e., 35%) and randomly placed circles and rectangles with reflectances ranging from 0.08 to 0.75. The panel was attached to a Plexiglas backing, which was supported by a metal frame. Centered in the panel were two circular apertures of diameter 2.5 cm, which were 6.5 cm apart, center-to-center. Immediately behind each aperture, flat against the back of the panel, was a disk of diameter 25 cm. Each disk had a circular metal backing with a circular paper printout attached to its surface. The paper showed an annulus whose reflectance ranged continuously from 0.06 to 0.80. A computer-controlled servo motor rotated each disk to adjust the part of the annulus (and hence the reflectance) seen through each aperture. Only a small area of the annulus was visible through the aperture at any time, and the range of reflectances visible due to the continuous reflectance gradient on the annulus was never greater than 6.3%, e.g., if the reflectance at one point of the aperture boundary was 0.2000, then the reflectance at opposite point was no greater than 0.2000 × 1.063 = 0.0216. As a result, the reflectance gradient was not perceptible. The same poster printer and type of paper were used to create the printouts on the panel and the disks.

We measured the reflectance of parts of the apparatus using a photometer (model LS-110; Konika Minolta, Tokyo, Japan) and a 99% diffuse reflectance standard (Spectralon SRS-99-020; Labsphere, North Sutton, NH). We measured the luminance L of a target surface location, and the luminance L_{99} of the reflectance standard placed at the same location. Both measurements were made under the illuminant used in the experiment (see below). We calculated the reflectance of the target location as $r = 0.99\, L/L_{99}$.

The apparatus and room were illuminated by overhead LED lights and a data projector. The LED lights were seven computer-controlled smart bulbs (Philips Hue, White and Color Ambiance E26, 1100 lumen; Signify N.V., Eindhoven, Netherlands) in an enclosed, spherical cloth lampshade of diameter 55 cm, located above, behind, and to the left of the observer (Fig. 3). The color and brightness of the LED lights were adjusted so that the 99% reflectance standard, placed at the center of the panel apparatus, had an xy chromaticity of (0.30, 0.32) and a luminance of 37 cd/m^2 as measured by a spectrophotometer (SpectraScan PR-655; JADAK, North Syracuse, NY).

The data projector was located above, behind, and to the right of the observer, and directed towards the panel apparatus on the table (model CP-EX252N; Hitachi Ltd, Tokyo, Japan; Fig. 3). We used the photometer and reflectance standard to measure the mapping from the greyscale RGB value displayed on the projector to the illuminance at the two apertures in the panel apparatus. The xy chromaticity of the projected light at the center of the panel apparatus was (0.30, 0.32). We displayed a two-tone image on the data projector in order to create different uniform illumination intensities on the left and right halves of the panel apparatus (Fig. 4(a)). (Specific illuminances are given below under Procedure.) The resulting illumination boundary extended beyond the panel apparatus, and created a shadow-like boundary on the table and surrounding furniture as well. We defocused the projector so that no pixelation was visible in the projected image, and to blur the lighting boundary, which as a result was penumbra-like and unlikely to be mistaken for a reflectance boundary.

Behind the table was a white bookshelf that displayed a range of objects, which we included to enrich the scene and provide visual information that observers could potentially use to estimate lighting conditions (Fig. 4(a)).

VR environment The observer viewed a scene in an Oculus Rift S headset (88° FOV, 1280 × 1440 pixel resolution per eye), driven by an NVIDIA GeForce GTX 1060 graphics card on a PC running Windows 10. The scene (Fig. 4(b)) was modelled after the room and apparatus in the physical environment, and rendered in Unity using the Built-in Render Pipeline (Unity Technologies, 2020 Version 2019.3.3f1). The virtual panel apparatus was carefully matched to the physical apparatus: the simulated size, position, and reflectance pattern of the virtual panel were the same as in the physical apparatus, as were the simulated size and positions of the two apertures. The panel apparatus was rendered as a Lambertian material (Unity material type Legacy/Diffuse). The rest of the virtual scene was also matched to the physical scene, though not as rigorously. The virtual room contained walls, furniture, and objects that we assigned approximately the same simulated size, position, chromaticity, and material properties as their physical counterparts. Size and position were replicated by assigning the virtual objects simulated sizes that matched the measured sizes of the physical objects. Chromaticity and material properties were replicated by adjusting the virtual material properties until the virtual objects appeared approximately the same as the physical objects. We also measured luminance at a coarse grid of locations and adjusted the virtual scene so that the luminances of the rendered RGB values were approximately proportional.

Located behind the observer, the virtual lighting consisted of a point light source and a directional light source (Unity light types Point and Directional, with the background 'skybox' set to black). A virtual occluding surface behind the observer blocked the directional light from illuminating half of the scene, which produced a lighting boundary down the middle of the panel apparatus and surrounding furniture, as in the physical environment. The occluding surface could be moved to the left or right, depending on whether the brighter side of the illumination boundary was to be on the left or right side of the apparatus. Due to the limited luminance range of the VR headset, luminances were lower in the VR environment than in the physical environment. (Specific values are given under Procedure.) However, luminances on the virtual panel apparatus were proportional to the corresponding luminances on the physical apparatus. We also configured the virtual lighting so that the luminance at several locations in the scene, on the walls surrounding the panel apparatus, was proportional to the luminances in the physical scene, to within approximately 10% error.

We calibrated the VR headset to ensure that the luminances displayed were proportional to rendered achromatic RGB values (i.e., the three integers in the range 0-255 that represent each pixel). Unity's Post-Processing Stack includes a feature called the 'LUT Texture' that passes rendered RGB values through a nonlinearity specified by the user. We used the LS-110 photometer to measure the nonlinear mapping from achromatic RGB value to physical luminance without the LUT Texture mechanism. We then used the LUT Texture to pass rendered achromatic RGB values through a compensating nonlinearity, with the result that physical luminance was proportional to the rendered achromatic RGB values. We describe this calibration procedure more thoroughly in a separate publication (Murray et al., 2022).

Flat-panel environment Observers viewed the same virtual panel apparatus as in the VR environment, but displayed on a flat-panel LCD monitor (Dell UltraSharp 27" U2719D; Dell, Round Rock, TX; Fig. 4(c)) in a dark room instead of a VR headset. The physical monitor was placed at a viewing distance of 71 cm, where its active area (59.5 cm × 43.5 cm) subtended the same horizontal visual angle as the panels in the physical and VR environments. The stimulus was rendered by the same Unity program used in the VR environment, with a virtual camera positioned so that the panel apparatus filled the full width of the monitor. We measured the monitor's gamma function using the LS-110 photometer, and linearized the mapping from achromatic RGB to luminance using the same post-processing method as in the VR environment. The luminances of the various regions of the panel were proportional to the corresponding luminances in the physical and VR conditions.

Procedure

Before starting the experiment, the observer made a few practice lightness matches in the physical condition (see below for details). This familiarized them with the task, and pro-

viding instructions while they used the physical apparatus made it easier to convey that the goal was to match the perceived shade of grey paper visible through the apertures (i.e., lightness), and not some other stimulus property such as illumination or luminance. Performance in matching tasks can be highly dependent on instructions (Blakeslee et al., 2008), so for consistency the experimenter read the instructions from a prepared script. After the instructions and practice trials, the observer ran in the physical, VR, and flat-panel conditions in a randomly chosen order. Each observer completed the experiment on a single day, with breaks of at least 20 min between the three conditions.

Physical condition The observer sat at a table supporting the experimental apparatus, and their head position was stabilized by a chinrest. The panel apparatus described above was approximately at eye level.

On each trial, one aperture (left or right) was randomly chosen as the reference aperture, and the other was the match aperture. The illuminance on the half of the apparatus containing the reference aperture was set to 116 lx using the data projector. The illuminance on the half containing the match aperture was randomly set to one of five values: 145, 181, 227, 283, or 354 lx, also using the data projector. The reflectance at the reference aperture was randomly set to one of three values: 0.18, 0.39, or 0.55. The reflectance at the match aperture was initialized to a random value within the range displayable on the apparatus (0.06 to 0.75). A red dot, 2.5 cm in diameter, was shown 10 cm below the chosen match aperture for one second using the data projector, to indicate which was the match aperture on the current trial. A short beep then indicated the start of the trial. The observer adjusted the reflectance visible through the match aperture by swiping up or down on a trackpad. A quick stroke made a large adjustment in the match reflectance, and a slow stroke made a fine adjustment. The observer was instructed to adjust the material visible through the match aperture so that it appeared to be the same shade of grey paper as the material visible through the reference aperture. The observer had unlimited time to make this setting, and pressed the spacebar on a keyboard to indicate that they had completed the match. A short beep acknowledged their response, and then the next trial began.

There were three reference reflectances and five match illuminances (given above). Each reflectance-illuminance combination was repeated five times with the left aperture as the reference, and five times with the right as the reference, for a total of $3 \times 5 \times (5 + 5) = 150$ trials.

VR condition This condition was largely the same as the physical condition, except that the observer viewed a virtual scene in a VR headset. Furthermore, we did not use a chinrest in this condition, in order to avoid contact between the headset and chinrest. Instead, a green cube was positioned at the intended eye position in the virtual scene. Participants were instructed to adjust their head position until the cube was no longer visible, ensuring a consistent viewpoint. If participants moved their head too far back, they could see the green cube again, prompting them to readjust their head position until the cube disappeared. This arrangement encouraged the observer to keep their head at a fixed viewing position, while still allowing small head movements, much as with the chinrest in the physical condition.

When the observer's head was located at the position of the green cube, the virtual camera was a simulated distance of approximately 71 cm from the panel apparatus, which was the same viewing distance as in the physical condition. The simulated reference reflectances were also the same as in the physical condition. The reference and match apertures in the apparatus were assigned uniform reflectances, instead of replicating the small and imperceptible reflectance gradient at each aperture in the physical apparatus. The reference and match illuminances were lower because of the limited luminance range of the headset, but they were proportional to the values in the physical condition. The simulated reference illuminance was 51 lx, and the simulated match illuminances were 63, 79, 99, 125, and 155 lx. The observer used two joysticks on VR controllers to adjust the match reflectance. The joystick on the left controller changed the reflectance quickly, and the one on the right controller changed it more slowly. The observer pressed a button on the right controller to indicate that they had completed the reflectance match.

Flat-panel condition This condition was largely the same as the physical condition, except that the observer viewed a rendering of the panel apparatus on a flat-panel LCD monitor. Only the panel apparatus, and no other element of the scene, was shown on the monitor, as the goal of this condition was to imitate a typical psychophysical experiment where only the primary stimulus elements are displayed, e.g., Allred et al. (2012); Toscani et al. (2016); Anderson and Winawer (2005). The monitor was placed on a table in a dark and otherwise empty testing room. The bookcase and objects that were behind the testing apparatus in the physical condition were not present in this condition. Head position was stabilized by a chinrest. The simulated reference reflectances were the same as in the other conditions. The simulated illuminances were higher than but proportional to those in the other two conditions: the simulated reference illuminance was 286 lx, and simulated match illuminances were 358, 447, 559, 699 and 874 lx. The observer adjusted the match reflectance using two joysticks on a gamepad, in a manner similar to the VR condition, and pressed a button on the gamepad to indicate that they had completed the match.

Results

Figure 5 shows two typical observers' match settings as a function of illuminance at the match patch. (Results for the remaining observers are shown in Appendix B.) The straight lines are sum-of-squares fits, constrained to pass through the points representing the reference stimuli, which are indicated by open circles. As explained above, the observer's Thouless ratio is given by one plus the slope of the fitted line, and this Thouless ratio is shown at the right of each line. In all three environment conditions, observers' match settings were well fit by straight lines, and match reflectance was only moderately affected by illuminance. (Code that produces this figure, as well as all other figures and analyses reported in the Results and Discussion sections, is provided as Supporting Information.)

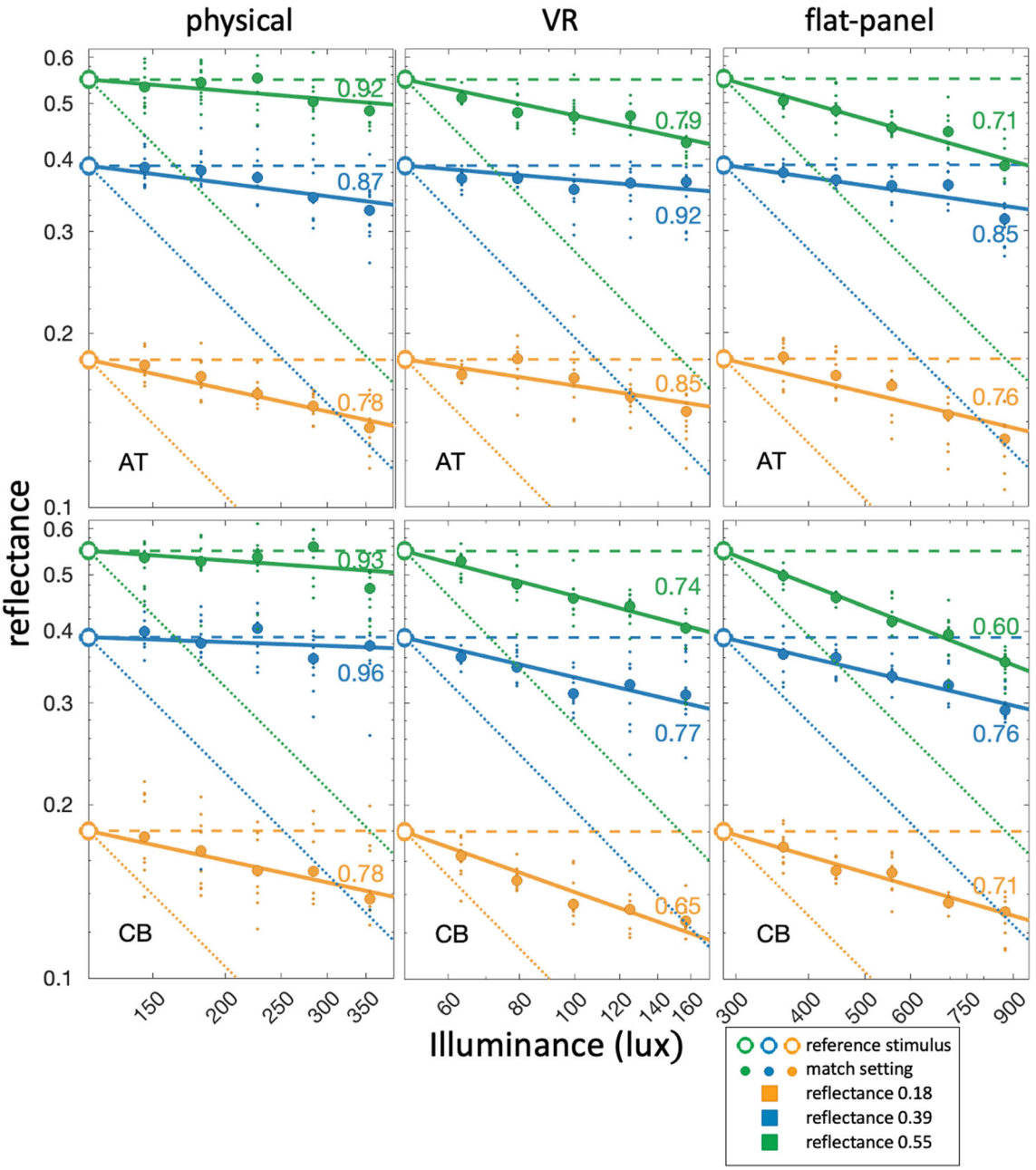

Fig. 5 Results for two typical observers in the lightness matching task. Small dots represent reflectance matches on individual trials, and large filled dots represent the median match across trials. Straight lines are sum-of-squares fits, constrained to pass through the open circles that represent the reference stimuli. The three colors in each panel show data for the three reference reflectances. The horizontal dashed lines with slope zero represent perfect constancy (τ=1), and the diagonal dotted lines with slope -1 represent luminance matching (τ=0)

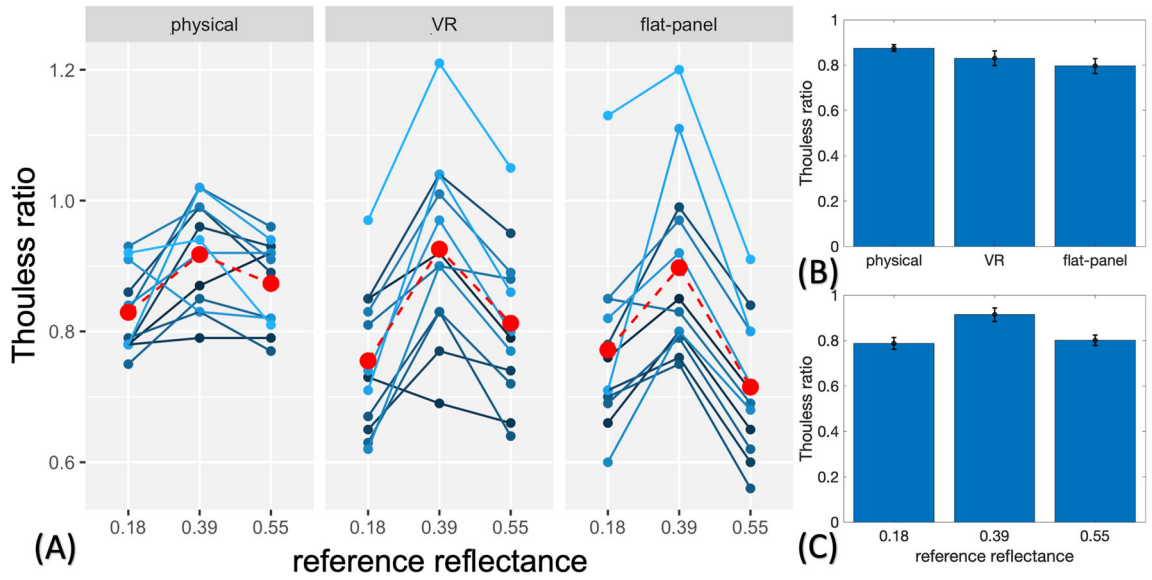

Fig. 6 **a** Thouless ratios as a function of environment and reference reflectance. Each triad of blue dots connected by straight lines represents a single observer, and red dots represent means across observers. **b** Mean Thouless ratios by environment. **c** Mean Thouless ratios by reference reflectance

Figure 6(a) shows Thouless ratios for all 12 observers in all conditions. Each set of three small blue dots connected by straight lines represents an individual observer, and the larger red dots represent mean Thouless ratios across observers. The mean Thouless ratio was 0.87 in the physical condition, 0.83 in the VR condition, and 0.79 in the flat-panel condition. Table 1 reports mean Thouless ratios in all conditions, as well as bootstrapped standard errors and 95% confidence intervals.

Figure 6(a) suggests that variability across observers was larger in the VR and flat-panel conditions than in the physical condition. A Levene's test for homogeneity of variance confirmed unequal variance across the three conditions ($F(2, 105) = 3.487$, $p < 0.05$). Post hoc pairwise comparisons showed that there was a significant difference in variability between the physical and flat-panel conditions ($F(1, 70) = 5.38$, $p < 0.05$), and between the physical and VR conditions ($F(1, 70) = 7.49$, $p < 0.05$) conditions, but not between the VR and flat-panel conditions ($F(1, 70) = 0.01$, $p > 0.05$).

Figure 6(a) also suggests that the high variability of Thouless ratios in the VR and flat-panel conditions was largely driven by individual differences, with observers tending to consistently show high or low Thouless ratios across reference reflectances and display types. To investigate this hypothesis, we plotted Thouless ratios in the flat-panel condition against the corresponding Thouless ratios in the VR condition, for each participant and reference reflectance (Fig. 7). There was a clear linear relationship between Thouless ratios for the two display conditions. A correlation test using Bonferroni corrected significance level of $p = 0.05/3 = 0.017$ showed significant Pearson correlations between Thouless ratios in the VR and flat-panel conditions for all three reference reflectances: reference reflectance 0.18 ($r(34) = 0.85$, $p < .017$, 95% CI [0.54, 0.96]), reference reflectance 0.39 ($r(34) = 0.91$, $p < .017$, 95% CI [0.71,

Table 1 Mean Thouless ratios by condition, with bootstrapped standard errors and confidence intervals

		Mean	Standard error	95% confidence interval
Environment	Physical	0.873	0.015	0.844 – 0.901
	VR	0.831	0.032	0.772 – 0.900
	Flat-panel	0.795	0.033	0.736 – 0.864
Reference Reflectance	0.18	0.785	0.026	0.738 – 0.839
	0.39	0.914	0.030	0.857 – 0.974
	0.55	0.800	0.023	0.755 – 0.844

The mean for each environment is averaged over reference reflectances and participants, and the mean for each reference reflectance is averaged over environments and participants

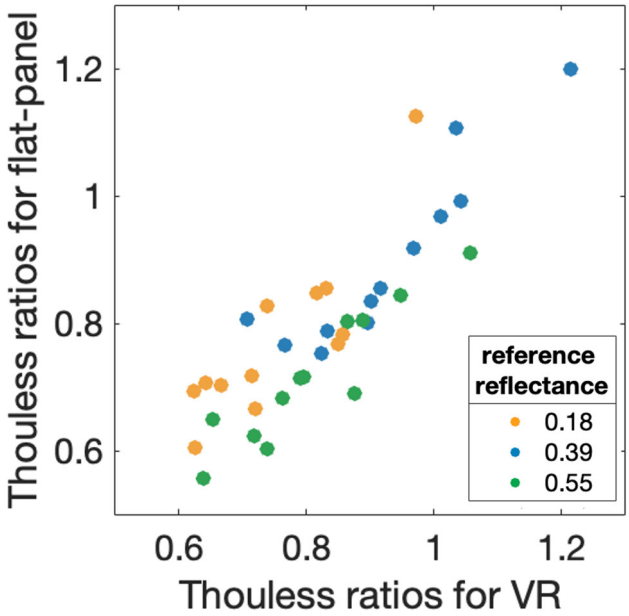

Fig. 7 Thouless ratios in the VR and flat-panel conditions are strongly correlated

0.98]), and reference reflectance 0.55 ($r(34) = 0.93$, $p < .017$, 95% CI [0.76, 0.98]).

As the Levene's test indicated non-homogeneous variances, we compared mean Thouless ratios across conditions using bootstrap tests in order to avoid making strong distributional assumptions. We created 10^6 bootstrapped replications of the experiment, where each replication was based on 12 observers randomly selected with replacement from the 12 observers in the experiment. For each bootstrapped replication we calculated the mean Thouless ratio across observers for each of the three environments (physical, VR, flat-panel) and three reference reflectances (0.18, 0.39, 0.55). (See Appendix C for further details of the bootstrapped significance tests.) A two-tailed comparison of the bootstrapped means showed no significant difference between the physical and VR conditions ($p = 0.12$). However, the physical condition had a significantly higher mean than the flat-panel condition ($p < 0.05$), and similarly, the VR condition had a significantly higher mean than the flat-panel condition ($p < 0.001$). Furthermore, the mean for reference reflectance 0.39 was significantly greater than the means for reference reflectances 0.18 and 0.55 ($p < 0.001$ in both cases), but the means for the latter two reference reflectances were not significantly different ($p = 0.41$).

These p values were not corrected for multiple comparisons. However, for all but one of the comparisons that achieved significance, the results were highly significant ($p < 0.001$), and even a conservative correction, such as a Bonferroni correction (criterion $p = 0.05/6 = 0.008$) , would still indicate significant differences. The one exception is the comparison of the mean Thouless ratio in the physical and flat-panel conditions ($p = 0.0137$), which a Bonferroni correction would render non-significant.

Figure 8 plots Thouless ratios against median response times, for each observer in each condition. The response method was slightly different in the three environments: a trackpad was used in the physical condition, and two different models of joysticks were used in the VR and flat-panel conditions. As a result, it is difficult to confidently compare response times across conditions, but nevertheless we note that response times were broadly similar: median response time was 6.6 s in the physical condition, 5.7 s in the VR condition, and 6.3 s in the flat-panel condition. Furthermore, a correlation test showed no significant Pearson correlation between Thouless ratio and median response time in the physical condition ($r(34) = -0.26$, $p > 0.05$, 95% CI [-0.54, 0.08]), the VR condition ($r(34) = 0.09$, $p > 0.05$, 95% CI [-0.24, 0.41]), or the flat-panel condition ($r(34) = -0.04$, $p > 0.05$, 95% CI [-0.36, 0.29]).

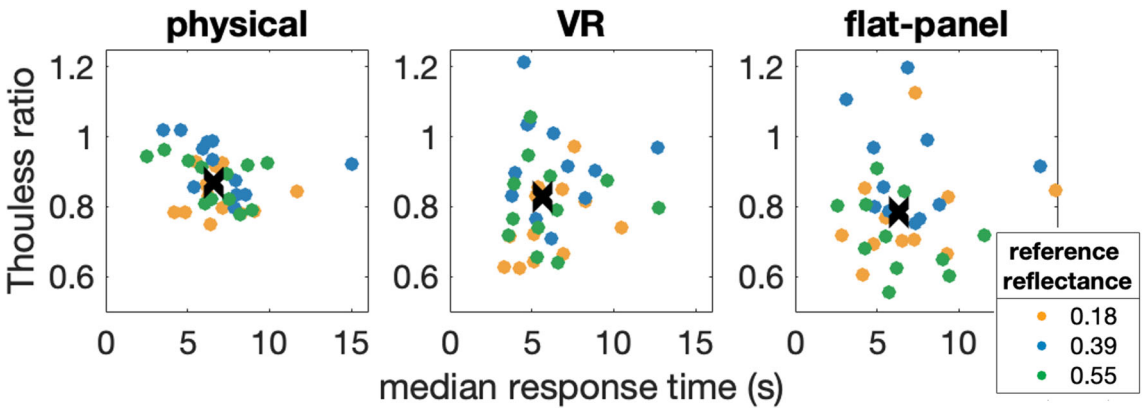

Fig. 8 Thouless ratio versus median response time. The black 'x' indicates the median response time and Thouless ratio for each environment. The standard errors for the medians were approximately the same size as the marker itself, and so are not displayed in this plot

Discussion

The goal of this study was to compare the levels of lightness constancy that observers achieve in real environments, in virtual environments, and on flat-panel displays. We found that in a simple 2D lightness matching task, observers had good constancy in all three environments, with Thouless ratios typically in the range 0.7 to 0.9. Constancy was not significantly different in the physical and VR environments. It was significantly better in the physical environment and VR environments than on the flat-panel monitor, but even there the effect size was not large, with mean Thouless ratios of 0.87 (physical) and 0.83 (VR) versus 0.79 (flat-panel).

The Thouless ratios reported in all three environments indicate good lightness constancy, compared to previous experiments with physical stimuli. For example, (Gilchrist, 2006 p. 31) re-examined data from Katz's pioneering lightness constancy experiments, and found Thouless ratios between 0.35 and 0.75. Gilchrist pointed out that Katz's apparatus provided somewhat impoverished lighting cues, resulting in lower Thouless ratios than in later studies. Patel et al. (2018) reported Thouless ratios with a mean of 0.82 in a condition with a physical lightness matching apparatus. Similar values are also typically reported in color constancy experiments that use a Thouless-like index to quantify constancy in rich, physical environments, such as Kraft and Brainard (1999), who reported a mean constancy index of 0.83 in a condition where a wide range of color and lighting cues were available.

Variability in virtual conditions An unexpected finding in the present experiment was that inter-subject variability in lightness constancy was significantly greater with VR and flat-panel displays than with the physical apparatus. We can only speculate on the reasons for this difference, but it may have occurred because the virtual environments did not provide completely realistic or consistent cues to lighting and surface properties. For example, the Lambertian material model that we used for the simulated apparatus is only an approximate representation of real matte materials (Guarnera et al., 2016), and furthermore commercial VR displays provide conflicting cues to depth and surface reflectance. An environment with inconsistent cues may amplify the effects of individual differences in how observers integrate multiple cues to judge surface properties (Westerman & Cribbin, 1998), or individual differences in observers' Bayesian priors for scene properties (Adams et al., 2004). We cannot be certain about this interpretation, though, and the higher variability of performance in virtual environments calls for further study.

A role for contrast polarity Another unexpected finding was that lightness constancy was significantly better for one reference reflectance (0.39) than for the other two (0.18, 0.55) (Fig. 6(c)). This difference was substantial: the Thouless ratio was 0.91 in the former condition, versus 0.79 and 0.80 in the latter two, which is larger than the difference between the real and flat-panel conditions discussed earlier (see Table 1). This effect might be explained by the fact that the reference reflectance 0.39 was just slightly higher than the reflectance of the background panel, which was 0.35. Human lightness matches cannot be completely explained in terms of contrast matching (Kraft & Brainard, 1999), but in this case, where the luminance of the reference patch was slightly higher than its surround, contrast polarity may have provided observers with a useful heuristic: they could use the constraint that the luminance of the match patch should also be slightly higher than its surround.

To illustrate this heuristic, here we calculate the match setting that observers would have made in the physical condition with reference reflectance 0.39, if their Thouless ratios had been similar to those for the other two reference reflectances. The grey background of the apparatus had reflectance 0.35, so a reference reflectance of 0.39 produced a positive-contrast reference patch. In the physical condition, the mean Thouless ratio for reference reflectances 0.18 and 0.55 was 0.80. Equation 8 shows that with reference reflectance 0.39, reference illuminance 116 lux, and match illuminance 354 lux (the highest value used in the experiment), the Thouless ratio 0.80 predicts a match reflectance of 0.31, which is a negative-contrast match setting on a background of reflectance 0.35. In fact, observers mostly made positive-contrast match settings in this condition, with a median match reflectance of 0.37. This simple calculation shows that if observers had the same Thouless ratio for reference reflectance 0.39 as for the other two reference reflectances, then under high match illuminances, they would have had to make a negative-contrast match to a positive-contrast reference stimulus. We speculate that observers were unwilling to make such a match, and instead gave the match patch a luminance slightly higher than its surround, resulting in a higher Thouless ratio. The other two reference reflectances were much higher or lower than the background reflectance, so in those cases the heuristic did not provide a useful constraint on observers' match settings.

We explored this hypothesis further by developing a simple model where an observer's match setting was determined by both an illumination discounting stage and a penalty on match settings that produced opposite contrast polarities at the reference and match patches. We did not numerically optimize the fit of this model to our data; rather, our goal was to show qualitatively that the contrast-polarity heuristic can produce inflated Thouless ratios for low-positive-contrast stimuli, as observed in our experiment with human observers. The Supporting Information includes a MATLAB implementation of the model.

The first component of the model used Eq. 8 to calculate an initial match reflectance r_{m*} that was not subject to the penalty on opposite-polarity matches described below.

This calculation assumed a Thouless ratio $\tau = 0.8$, which was the average value for human observers for reference reflectances 0.18 and 0.55. τ was therefore a free parameter of the model. To make this step more realistic as a model of human behavior, the model added a random perturbation to the estimate. The random component was a sample from a normal distribution with mean zero and standard deviation $\sigma_1 = 0.03$, which was the median standard deviation of human observers' match settings.

The second component of the model chose a match reflectance r_m by minimizing an objective function with two terms. The first term f_1 was the squared difference between r_m and the match reflectance r_{m*} described in the previous paragraph:

$$f_1(r_m) = (r_m - r_{m*})^2 \tag{4}$$

The second term f_2 penalized match settings r_m that gave the match patch a different local contrast polarity than the reference patch. That is, on trials where the reference reflectance r_{ref} was higher than the background reflectance r_{bg}, the model penalized match reflectance settings r_m that were lower than the background reflectance r_{bg}, and vice versa. The penalty term was

$$f_2(r_m) = F((r_{bg} - r_m)sgn(r_{ref} - r_{bg}), \alpha, \beta) \tag{5}$$

Here $F(x, \alpha, \beta)$ is the Weibull cumulative distribution function, and $\mathrm{sgn}(x)$ is the signum function:

$$\mathrm{sgn}(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ +1 & x > 0 \end{cases} \tag{6}$$

We plot the penalty function f_2 in Fig. 9(a). α is a scale parameter that determines the match reflectance at which the penalty function saturates; we used $\alpha = 0.05$. β is a shape parameter; we used $\beta = 2$. $F(x, \alpha, \beta)$ is nonzero only when $x > 0$, so the sgn factor causes the penalty function to penalize positive values of $r_m - r_{bg}$ when $r_{ref} < r_{bg}$, and to penalize negative values when $r_{ref} > r_{bg}$.

The complete objective function was

$$f(r_m) = f_1(r_m) + wf_2(r_m) \tag{7}$$

The weight w determined the maximum value of the contrast-polarity penalty, and we used $w = 0.1$. The model's match setting was the reflectance that minimized the objective function f. This value varied from trial to trial, because f_1 depended on the initial match reflectance r_{m*}, which was stochastic as explained above.

We computed 1000 match reflectance settings (simulating 1000 trials) for each combination of reference reflectance and match illuminance used in the experiment with human observers. Figure 9(b) shows the results. The model had a Thouless ratio of 0.8 for reference reflectances 0.18 and 0.55; the penalty term had little effect in these cases, where the reference reflectances were far above or below the background reflectance of 0.35. The model had a Thouless ratio of 0.88 for the reference reflectance 0.39; here the penalty term favored positive-contrast match settings, resulting in a higher Thouless ratio. Note that the penalty term was not a hard constraint: even with reference reflectance 0.39, the model sometimes produced match settings lower than the background reflectance (0.35), as was the case with human observers as well. These Thouless ratios are practically identical to those reported above for human observers, and overall these results show that the contrast-polarity heuristic is a

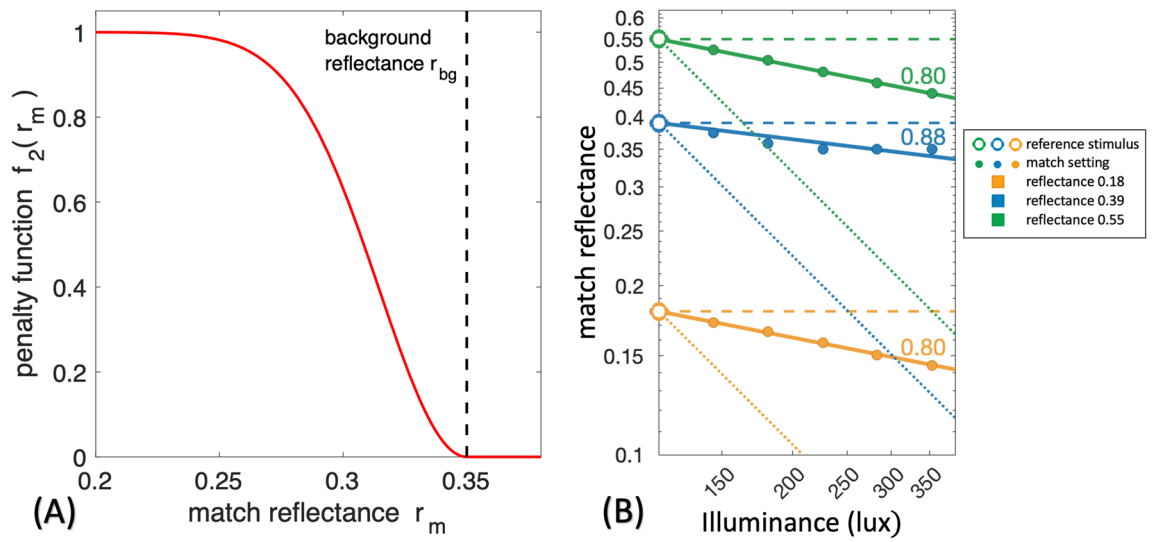

Fig. 9 (a) The penalty function f_2. (b) The model's median match settings and the corresponding Thouless ratios

plausible explanation of the higher Thouless ratios that we observed for the reference reflectance (0.39) that was close to the background reflectance.

The contrast-polarity model we have presented here is somewhat ad hoc, as it simply imposes a penalty on opposite-polarity matches. A more thorough and satisfactory account would motivate this penalty term, for example in terms of natural scene statistics. However, here our goal has been to go beyond speculation regarding why some reflectance matches were markedly higher than others and to show that, in principle, some form of contrast-polarity penalty can quantitatively account for this effect. The exact nature of this penalty and the conditions under which it is imposed are questions for further study.

Comparison to previous studies Blakeslee et al. (2008) also examined lightness judgements in real and VR environments. The goal of their study, however, was to examine a range of simultaneous contrast effects, not to measure lightness constancy, or to systematically compare performance in real and VR environments. Accordingly, their stimuli were simple, classic simultaneous contrast figures, and did not provide rich lighting cues. As a result, their observers reported that lightness judgements were effortful, conscious calculations, rather than spontaneous judgements of perceived surface reflectance. For these reasons, Blakeslee et al.'s results are not immediately comparable to ours, but it is still worth noting that their observers made similar lightness matches in real and VR environments, consistent with our findings.

How consistent is our main finding, namely that there was no significant difference in lightness constancy between real and VR environments, with the previous studies reviewed in the introduction? There may seem to be a discrepancy between our findings and Morgenstern et al. (2014), which found substantial differences between real and virtual environments. However, there are several possible reasons for these divergent findings. First, as noted earlier, Morgenstern et al. compared lightness constancy between two previous studies that were designed and run independently, and had many stimulus differences (Boyaci et al., 2003; Bloj et al., 2004). As a result, Morgenstern et al.'s comparison is not as reliable as a comparison between experiments with similar stimuli and tasks, as in the present study.

Second, Boyaci et al. (2003), which provided data for the virtual part of Morgenstern et al.'s comparison, used a stationary, custom-built stereoscopic display, not a VR headset. It may be that the immersive VR environment in the present study contributed to good lightness constancy.

Third, and we believe most importantly, the task in our study was fundamentally different from the one in Morgenstern et al. In our study, the task was to match the reflectance of two patches on a single frontoparallel surface, surrounded by the same background reflectance, and separated by a single shadow boundary. In the studies examined by Morgenstern et al., the task was to view an isolated reference patch in a complex scene, rotated to a new 3D orientation on each trial, and to choose a lightness match from a separate frontoparallel palette of match patches in a different part of the scene. It seems plausible that the latter task was intrinsically more demanding, as it required observers to compensate for the intensity of incident light at the 3D orientation of the reference patch (which changed from trial to trial), and to compare the resulting reflectance estimate to a palette at a different orientation and location. It may be that observers achieve good lightness constancy in VR environments on simple tasks like ours, but that more demanding tasks like those examined by Morgenstern et al. reveal shortcomings of virtual environments. This is a promising avenue for future work.

Conclusion In conclusion, we found similar levels of lightness constancy in a simple 2D lightness matching task in real and VR environments, and only slightly poorer constancy on a flat-panel display. Inter-observer variance, however, was substantially greater with VR and flat-panel displays than with the physical apparatus. This discrepancy should be considered when developing applications where realistic performance is critical, but overall our results show that VR can often be a flexible alternative to flat-panel displays, and a reasonable proxy for real environments.

Open practices statement

The data and supporting analysis code for all experiments are provided on OSF at https://doi.org/10.17605/OSF.IO/7EUYZ

Appendix A

We assume that all surfaces are Lambertian. Let the reference stimulus have reflectance r_1 and illuminance i_1, so following Eq. 1 for a Lambertian surface, the luminance is $\ell_1 = r_1 i_1/\pi$. Let the observer's match setting at the test stimulus be reflectance r_m, under illuminance i_m, and so have luminance $\ell_m = r_m i_m/\pi$. If the observer has no lightness constancy and simply matches the luminance of the reference and test stimuli, then their match setting r_0 satisfies $r_0 i_m = r_1 i_1$, or $r_0 = r_1 i_1/i_m$. If we substitute this expression for r_0 into Eq. 3 and solve for the match setting $\log r_m$, we find

$$\log r_m = (\tau - 1)(\log i_m - \log i_1) + \log r_1 \tag{8}$$

which is an affine function of $\log i_m$ with slope $m = \tau - 1$.

Appendix B

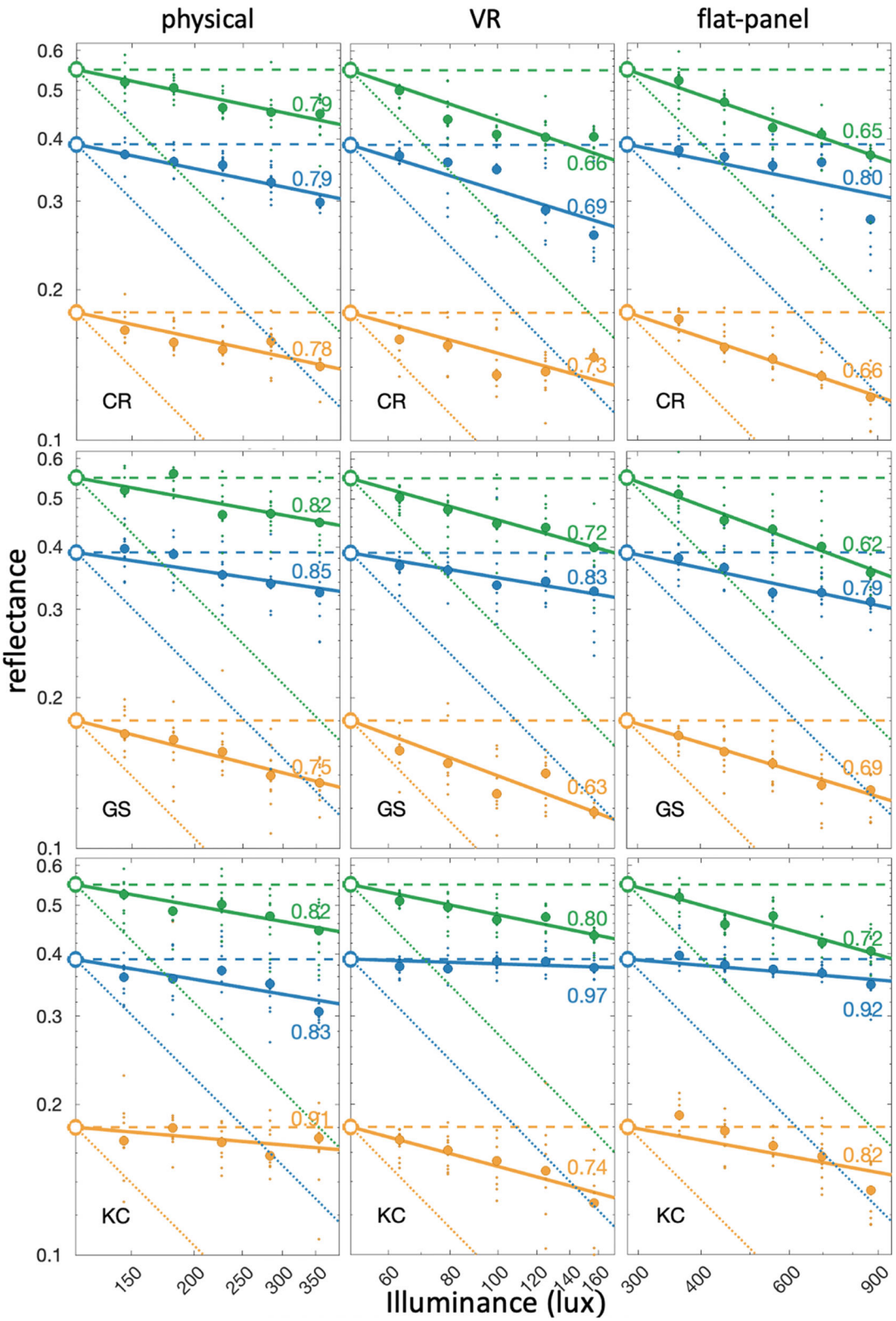

Fig. 10 Results for additional observers in the lightness matching task. See caption of Fig. 5 for details

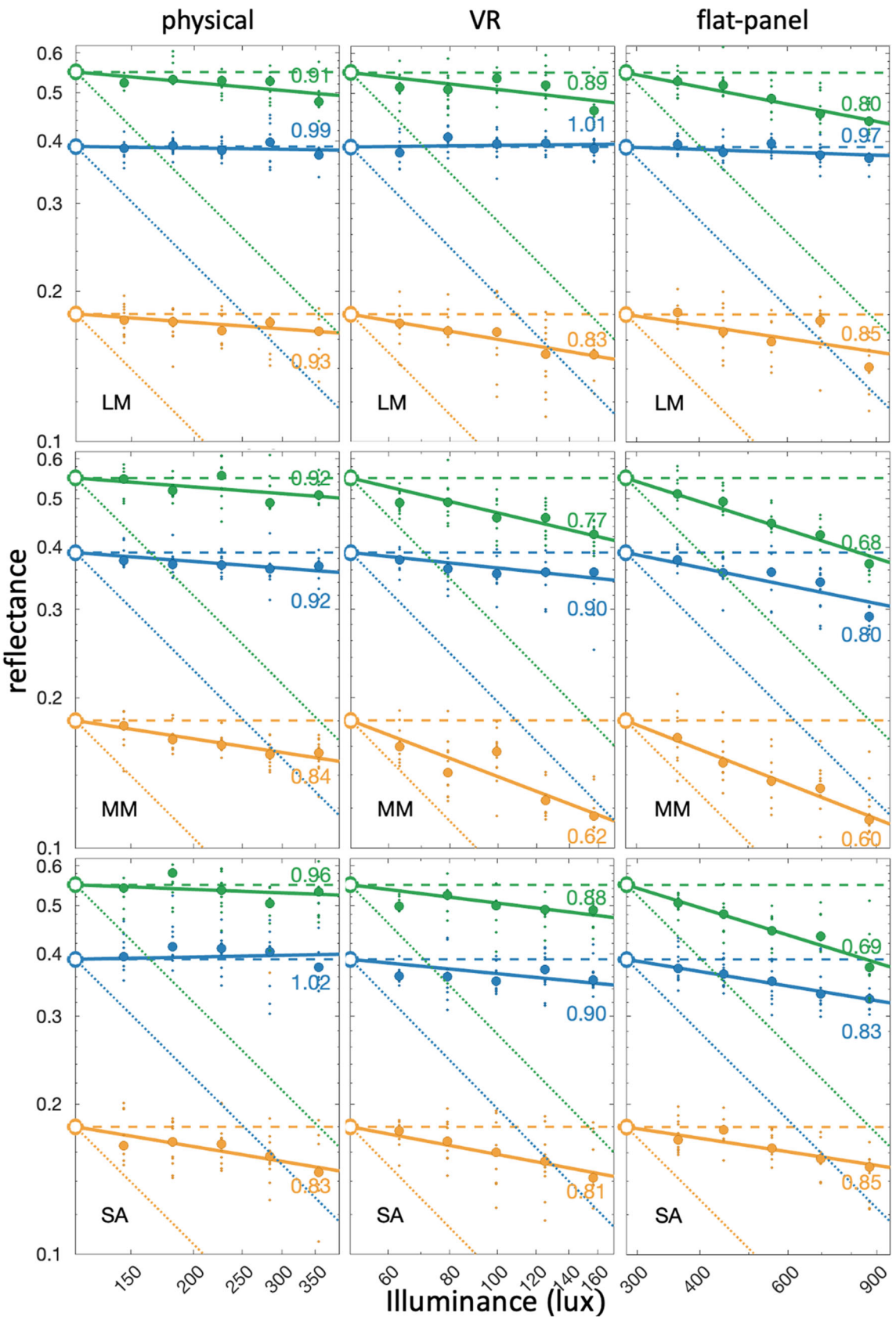

Fig. 11 Results for additional observers in the lightness matching task. See caption of Fig. 5 for details

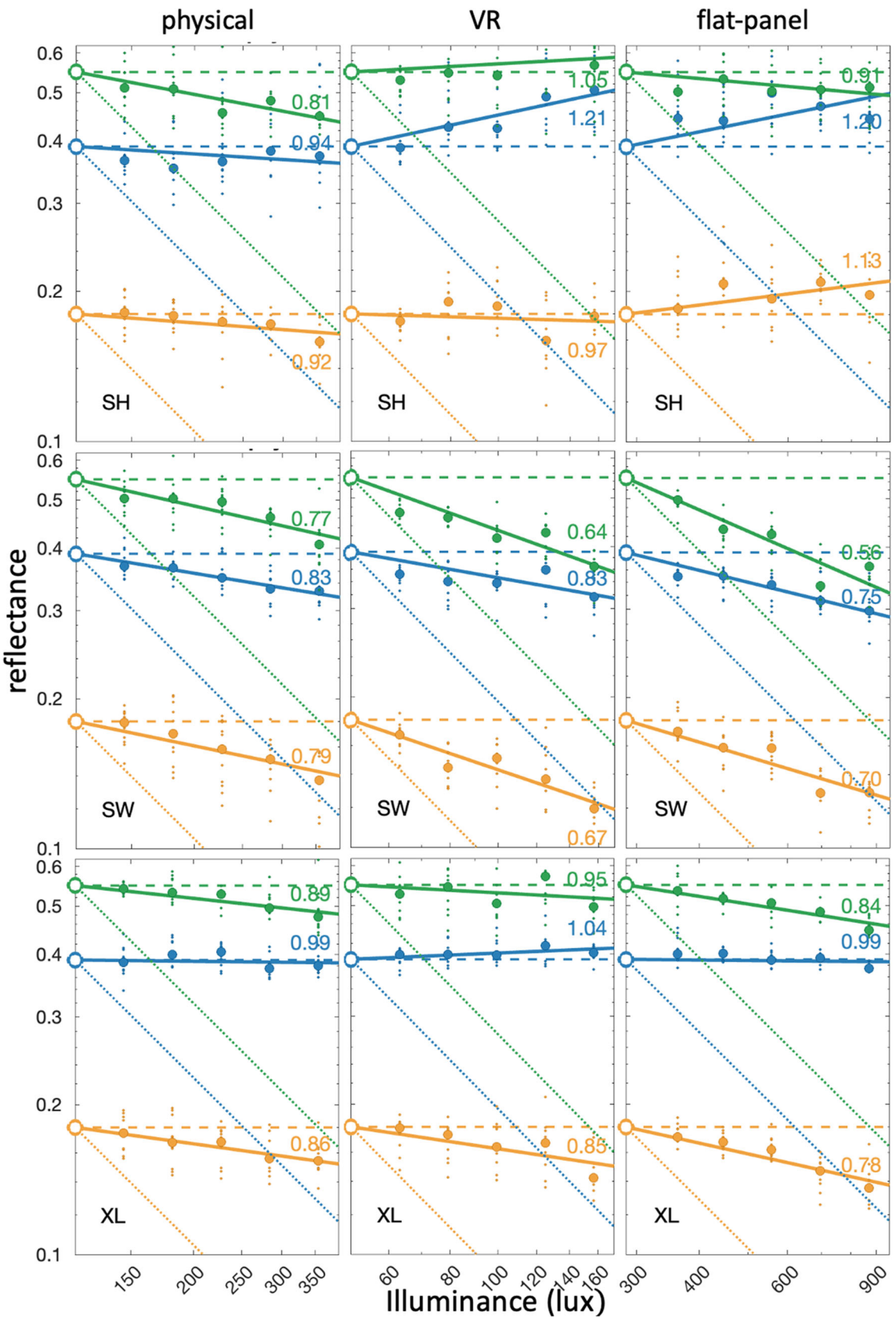

Fig. 12 Results for additional observers in the lightness matching task. See caption of Fig. 5 for details

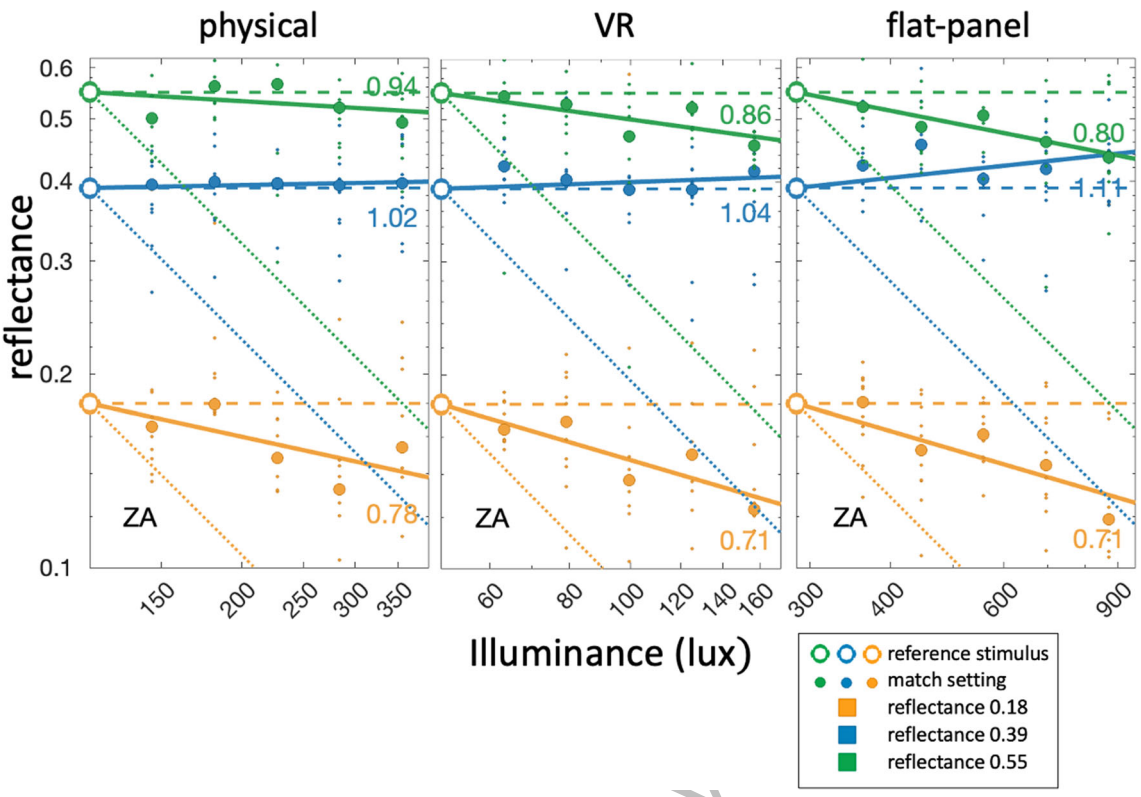

Fig. 13 Results for an additional observer in the lightness matching task. See caption of Fig. 5 for details

Appendix C

Here we provide details of the bootstrapped significance tests of mean Thouless ratios.

Figure 6 shows Thouless ratios for each display method, reference reflectance, and observer. This data can be represented as a $3 \times 3 \times 12$ matrix T_{ijk}, where each entry is the Thouless ratio for display method i (an integer from 1 to 3), reference reflectance j (also an integer from 1 to 3), and observer k (an integer from 1 to 12). The red dots in Fig. 6 show mean Thouless ratios across observers, which can be represented as a 3×3 matrix M_{ij}, where each element is the average of T_{ijk} for k from 1 to 12.

On each bootstrap iteration, we simulate a repetition of the experiment by creating a new $3 \times 3 \times 12$ matrix $T^{(b)}_{ijk}$ of Thouless ratios. We generate this matrix by choosing 12 observers with replacement from the 12 observers in the experiment. That is, each 3×3 slice of $T^{(b)}_{ijk}$ for a given value of k is a randomly chosen 3×3 slice of the original data T_{ijk}, representing the Thouless ratios of a single observer. ("With replacement" means that the observer's data may appear more than once in the resampled matrix $T^{(b)}_{ijk}$.) We then calculate the bootstrapped 3×3 matrix of means $M^{(b)}_{ij}$ by taking the average of $T^{(b)}_{ijk}$ over k from 1 to 12.

We repeat this sampling procedure $B = 10^6$ times, producing B simulated repetitions of the experiment, represented as $T^{(b)}_{ijk}$ and $M^{(b)}_{ij}$, where b ranges from 1 to B.

This resampled data forms the basis of the bootstrapped significance tests. For example, to test whether the mean Thouless ratio in condition $i = 1$, $j = 1$ (say, the physical condition and reference reflectance 0.18) is significantly greater than the Thouless ratio in condition $i = 1$, $j = 2$ (say, the physical condition and reference reflectance 0.39), we find the proportion of bootstrapped samples for which $M^{(b)}_{11}$ is greater than $M^{(b)}_{12}$. If this is true for at least 95% of the samples, then we conclude that the first mean is significantly greater than the second mean at a significance level of $p<0.05$.

For additional details of bootstrapping methods, see Efron and Tibshirani (1994). We provide MATLAB code that implements these significance tests as Supporting Information.

References

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience, 7*(10), 1057–1058.

Adelson, E. H. (2000). Lightness perception and lightness illusions. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 339–351). The MIT Press.

Allred, S. R., Radonjić, A., Gilchrist, A. L., & Brainard, D. H. (2012). Lightness perception in high dynamic range images: Local and remote luminance effects. *Journal of Vision, 12*(2), 7–7.

Anderson, B. L., & Winawer, J. (2005). Image segmentation and lightness perception. *Nature, 434*(7029), 79–83.

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

Barron, J. T., & Malik, J. (2015). Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(8), 1670–1687.

Belhumeur, P. N., Kriegman, D. J., & Yuille, A. L. (1999). The bas-relief ambiguity. *International Journal of Computer Vision, 35*(1), 33–44.

Berkman, M. I., & Akan, E. (2019). Presence and immersion in virtual reality. In N. Lee (Ed.), *Encyclopedia of computer graphics and games.* Springer.

Blakeslee, B., Reetz, D., & McCourt, M. (2008). Coming to terms with lightness and brightness: Effects of stimulus configuration and instructions on brightness and lightness judgments. *Journal of Vision, 8*(11), 1–18.

Bloj, M., Kersten, D., & Hurlbert, A. (1999). Perception of three-dimensional shape influences colour perception through mutual illumination. *Nature, 402*(6764), 877–879.

Bloj, M., Ripamonti, C., Mitha, K., Hauck, R., Greenwald, S., & Brainard, D. H. (2004). An equivalent illuminant model for the effect of surface slant on perceived lightness. *Journal of Vision, 4*(9), 6.

Boyaci, H., Maloney, L. T., & Hersh, S. (2003). The effect of perceived surface orientation on perceived surface albedo in binocularly viewed scenes. *Journal of Vision, 3*(8), 541–553.

Brainard, D. H. (1998). Color constancy in the nearly natural image. 2. Achromatic loci. *Journal of the Optical Society of America A, 15*, 307–325.

Brainard, D. H., & Maloney, L. T. (2011). Surface color perception and equivalent illumination models. *Journal of Vision, 11*(5), 1–1.

Brunswik, E. (1928). Zur Entwicklung der Albedowahrnehmung. *Zeitschrift für Psychologie, 109*, 40–115.

Creem-Regehr, S. H., Stefanucci, J. K., & Bodenheimer, B. (2022). Perceiving distance in virtual reality: Theoretical insights from contemporary technologies. *Philosophical Transactions of the Royal Society B, 378*(20210456), 1–12.

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. Chapman and Hall/CRC Press.

Elor, A., & Kurniawan, S. (2020). The ultimate display for physical rehabilitation: a bridging review on immersive virtual reality. *Frontiers in Virtual Reality, 1*(585993), 1–17.

Fairchild, M. D. (2013). *Color appearance models* (3rd ed.). West Sussex, UK: John Wiley & Sons Ltd.

Fechner, G. (1860/1966). Elements of psychophysics. New York, NY: Holt, Rinehart and Winston.

Gil Rodríguez, R., Bayer, F., Toscani, M., Guarnera, D., Guarnera, G. C., & Gegenfurtner, K. R. (2022). Colour calibration of a head mounted display for colour vision research using virtual reality. *SN Computer Science, 3*(1), 1–10.

Gilchrist, A. (2006). *Seeing black and white*. New York: Oxford University Press.

Greengard, S. (2019). *Virtual reality*. The MIT Press.

Guarnera, D., Guarnera, G. C., Ghosh, A., Denk, C., & Glencross, M. (2016). Brdf representation and acquisition. *Computer Graphics Forum, 35*, 625–650.

Hartle, B., & Wilcox, L. M. (2022). Stereoscopic depth constancy for physical objects and their virtual counterparts. *Journal of Vision, 22*(4), 1–19.

Hibbard, P. B. (2023). Virtual reality for vision science. In *Topics in behavioral neurosciences* (pp. 1–29). Springer.

Jung, S., & Lindeman, R. W. (2021). Does realism improve presence in VR? Suggesting a model and metric for VR experience evaluation. *Frontiers in Virtual Reality, 2*(693327), 1–7.

Kimura, K., Reichert, J. F., Olson, A., Pouya, O. R., Wang, X., Moussavi, Z., & Kelly, D. M. (2017). Orientation in virtual reality does not fully measure up to the real-world. *Scientific Reports, 7*(1), 1–8.

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.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research, 35*(3), 389–412.

Li, Z., Shafiei, M., Ramamoorthi, R., Sunkavalli, K., & Chandraker, M. (2020). Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2475–2484

Linowes, J. (2020). Unity 2020 virtual reality projects: Learn vr development by building immersive applications and games with unity 2019.4 and later versions. Packt Publishing Ltd.

Marantz, A. (2016). Studio 360. The New Yorker, April 25, 2016, pp. 86–94

McCluney, R. (1994). Introduction to radiometry and photometry. Artech House, Inc.

Morgenstern, Y., Geisler, W. S., & Murray, R. F. (2014). Human vision is attuned to the diffuseness of natural light. *Journal of Vision, 14*(9), 1–18.

Murray, R. F., Patel, K. Y., & Wiedenmann, E. S. (2022). Luminance calibration of virtual reality displays in unity. *Journal of Vision, 22*(13):1 , 1–9.

Murray, R. F. (2021). Lightness perception in complex scenes. *Annual Review of Vision Science, 7*, 417–436.

Ong, T., Wilczewski, H., Soni, H., Nisbet, Q., Paige, S. R., Barrera, J., & Bunnell, B. (2022). The symbiosis of virtual reality exposure therapy and telemental health: A review. *Frontiers in Virtual Reality, 3*(848066), 1–11.

Ostrovsky, Y., Cavanagh, P., & Sinha, P. (2005). Perceiving illumination inconsistencies in scenes. *Perception, 34*(11), 1301–1314.

Patel, K., Munasinghe, A., & Murray, R. (2018). Lightness matching and perceptual similarity. *Journal of Vision, 18*(5), 1–18.

Radonjić, A., Pearce, B., Aston, S., Krieger, A., Dubin, H., Cottaris, N. P., & Hurlbert, A. C. (2016). Illumination discrimination in real and simulated scenes. *Journal of Vision, 16*(11), 1–18.

Rzepka, A. M., Hussey, K. J., Maltz, M. V., Babin, K., Wilcox, L. M., & Culham, J. C. (2023). Familiar size affects perception differently in virtual reality and the real world. *Philosophical Transactions of the Royal Society B, 378*(1869), 1–14.

Sanchez-Vives, M. V., & Slater, M. (2005). From presence to consciousness through virtual reality. *Nature Reviews Neuroscience, 6*(4), 332–339.

Scarfe, P., & Glennerster, A. (2015). Using high-fidelity virtual reality to study perception in freely moving observers. *Journal of Vision, 15*(9), 3–3.

Scarfe, P., & Glennerster, A. (2019). The science behind virtual reality displays. *Annual Review of Vision Science, 5*, 529–547.

Slater, M., Lotto, B., Arnold, M. M., & Sánchez-Vives, M. V. (2009). How we experience immersive virtual environments: The concept of presence and its measurement. *Anuario de Psicología, 40*, 193–210.

Slater, M. (2018). Immersion and the illusion of presence in virtual reality. *British Journal of Psychology, 109*(3), 431–433.

Thouless, R. (1931). Phenomenal regression to the real object. *British Journal of Psychology, 21*(4), 339–359.

Toscani, M., Zdravković, S., & Gegenfurtner, K. R. (2016). Lightness perception for surfaces moving through different illumination levels. *Journal of Vision, 16*(15), 21–21.

Unity Technologies. (2020). Unity, version 2019.3.3f1. Retrieved from https://docs.unity3d.com/2020.1/Documentation/Manual/UnityManual.html

Westerman, S. J., & Cribbin, T. (1998). Individual differences in the use of depth cues: Implications for computer- and video-based tasks. *Acta Psychologica, 99*, 293–310.

Wilder, J. D., Adams, W. J., & Murray, R. F. (2019). Shape from shading under inconsistent illumination. *Journal of Vision, 19*(6), 2.

Xie, B., Alghofalli, R., Zhang, Y., Jiang, Y., Lobo, F. D., Li, C., & Yu, L.-F. (2021). A review on virtual reality skill training applications. *Frontiers in Virtual Reality, 2*(645153), 1–19.

Xiong, J., Hsiang, E.-L., He, Z., Zhan, T., & Wu, S.-T. (2021). Augmented reality and virtual reality displays: Emerging technologies and future perspectives. *Light: Science & Applications, 10*(216), 1–30.

Zhan, T., Yin, K., Xiong, J., He, Z., & Wu, S.-T. (2020). Augmented reality and virtual reality displays: Perspectives and challenges. *iScience, 23*(101397), 1–13.