

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

2 Khushbu Y. Patel¹, Laurie M. Wilcox¹, Laurence T. Maloney², Krista A. Ehinger³,
3 Jaykishan Y. Patel¹, Emma Wiedenmann^{1,4}, and Richard F. Murray¹

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

5 ²Department of Psychology, New York University

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

7 ⁴Department of Psychology, Carl Von Ossietzky Universität Oldenburg

8

Abstract

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

1

2

Introduction

3 In recent years there have been rapid advances in virtual reality (VR) technology
4 (Greengard, 2019; Xiong, Hsiang, He, Zhan, & Wu, 2021; Zhan, Yin, Xiong, He, & Wu,
5 2020), and VR has found a wide range of recreational and professional uses, including
6 gaming (Linowes, 2020), cinema (Marantz, 2016), training (Xie et al., 2021), physical
7 rehabilitation (Elor & Kurniawan, 2020), psychological therapy (Ong et al., 2022), and
8 vision research (Hibbard, 2023; Scarfe & Glennerster, 2015, 2019). In many applications,
9 key goals for VR design include *immersion*, the ability of the VR environment to
10 provide rich, interactive stimuli, and *presence*, the viewer's subjective sense of being
11 fluidly engaged in the environment (Berkman & Akan, 2019; Sanchez-Vives & Slater,
12 2005; Slater, Lotto, Arnold, & Sánchez-Vives, 2009). In some applications, an
13 additional goal is *realism*, meaning that people and things should appear to the viewer
14 in VR just as they would appear in the real world (Jung & Lindeman, 2021). Realism is
15 important, for example, in some kinds of training, where the participant learns skills
16 that must transfer to the real world, and it is also important in vision research, where
17 VR is often used to show precisely controlled, computer-generated stimuli that are
18 stand-ins for real objects and scenes. Interestingly, the available evidence suggests that
19 these three goals are only loosely related, e.g., a high degree of realism may not be a
20 prerequisite for strong presence (Jung & Lindeman, 2021; Slater, 2018)

21 Here we evaluate a specific but fundamental aspect of realism in VR, namely the
22 extent to which viewers perceive black, white, and grey surface colours in virtual
23 environments with the same accuracy as in real environments. We report achromatic
24 colour matching experiments with a physical apparatus that uses real lights and
25 surfaces, a VR environment designed to simulate the physical apparatus, and a
26 flat-panel monitor that shows a rendered image of the physical apparatus. Our goal is
27 to quantify and compare perception of achromatic surface colour in these three
28 environments, in order to evaluate the extent to which VR and flat-panel displays can
29 substitute for real lights and surfaces in applications where realism is important.

30 Lightness perception

31 Visual processing begins with stimulation of the 2D array of photoreceptors in the
32 retina, but the retinal image provides only indirect and ambiguous information about
33 useful properties of things we see (Belhumeur, Kriegman, & Yuille, 1999; Murray,
34 2021). For example, it is often useful to know the *surface colour* of things in the
35 environment: this apple is green, that fox is light grey, and so on. The retinal image,
36 though, depends both on the intrinsic surface colours of objects (i.e., the proportion of
37 light they reflect at various wavelengths), and also on the spectrum and intensity of

illumination in the environment. In order to perceive surface colour accurately, the human visual system must disentangle the effects of surface colour and lighting. This is a difficult computational problem that is still not well understood (Barron & Malik, 2015; Li, Shafiei, Ramamoorthi, Sunkavalli, & Chandraker, 2020), but remarkably, the human visual system usually accomplishes it with a high degree of accuracy (Brainard & Maloney, 2011; Murray, 2021). This ability to perceive surface colour accurately in a wide range of lighting conditions and environments is called *colour constancy*.

In work on surface colour perception and colour constancy, researchers sometimes simplify the problem under study by limiting stimuli to achromatic colours: 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 surface¹. 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 defined² as perceived reflectance (Gilchrist, 2006). For example, in a printed copy of the snake illusion (Figure 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).

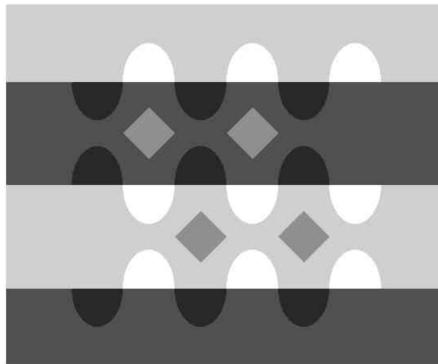


Figure 1. The snake illusion (Adelson, 2000). The four diamonds are physically identical, but the top two appear lighter than the bottom two.

Lightness perception raises many of the same problems as colour perception more generally. When we view a grey surface, the retinal light intensity depends on both the

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.

- 1 surface's reflectance and the amount of light incident on the surface. For a matte (i.e.,
 2 Lambertian) surface, the luminance (l) seen by an observer is proportional to the
 3 surface's reflectance (r) and to the illuminance (i) incident on the surface.

$$l = \frac{ir}{\pi} \quad (1)$$

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

9 Asymmetric lightness matching

10 A common experimental method for studying lightness perception is asymmetric
 11 lightness matching. In this design, the observer is shown a grey reference patch under
 12 one illuminant, and an adjustable match patch under another illuminant³ (Figure 2(a)).
 13 The observer's task is to adjust the reflectance of the match patch so that it appears to
 14 be the same as the reflectance of the reference patch. An observer who has perfect
 15 lightness constancy will make this match setting accurately, even though the two
 16 patches are shown under different illuminants. One who has poor lightness constancy,
 17 and for example matches the luminance of the two patches instead of the reflectance,
 18 will make inaccurate match settings. Thus the observer's match settings can be used to
 19 evaluate their lightness constancy under various experimental conditions.

20 One metric for quantifying lightness constancy is the *Brunswik ratio* (Brunswik,
 21 1928). Suppose that in an asymmetric lightness matching task, the correct reflectance
 22 match setting is r_1 (indicating perfect constancy), and the reflectance setting that
 23 would result from luminance matching is r_0 (indicating a complete lack of constancy).
 24 If the observer's actual match setting is r_m , then their Brunswik ratio is

$$\beta = \frac{r_m - r_0}{r_1 - r_0} \quad (2)$$

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

29 The *Thouless ratio* is a modification of the Brunswik ratio that takes into account
 30 the perceptual scaling of lightness (Thouless, 1931). Many studies have shown that

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

1 lightness is a compressive function of reflectance, e.g., the perceived difference between
 2 reflectances 0.10 and 0.15 is much larger than the perceived difference between 0.60 and
 3 0.65. In the Munsell colour system and CIELAB colour space, for example, which are
 4 attempts at perceptually uniform spaces where perceived colour differences are
 5 proportional to colour coordinate differences, the lightness dimension is approximately
 6 proportional to the cube root (a compressive function) of the corresponding physical
 7 dimension (Fairchild, 2013, pp. 80, 100). Similarly, the Weber-Fechner law states that a
 8 perceived quantity is a logarithmic function of the corresponding physical quantity
 9 (Fechner, 1860/1966) – again, a compressive mapping. The Thouless ratio uses a
 10 logarithmic transform of reflectance, and is defined as

$$\tau = \frac{\log r_m - \log r_0}{\log r_1 - \log r_0} \quad (3)$$

11 Like the Brunswik ratio, the Thouless ratio indicates where the observer's match setting
 12 lies between two theoretically defined reference points, corresponding to perfect
 13 constancy and no constancy. It is also similar to the 'colour constancy index' that
 14 indicates where an observer's colour match setting lies relative to two reference points
 15 in an approximately perceptually uniform colour space (Arend, Reeves, Schirillo, &
 16 Goldstein, 1991; Brainard, 1998). The values of Thouless ratios found in lightness
 17 matching experiments depend strongly on stimulus properties such as the complexity of
 18 the scene, but typical values in rich, naturalistic scenes are around 0.8 (Patel,
 19 Munasinghe, & Murray, 2018).

20 In an asymmetric lightness matching task such as the one we report below, the
 21 observer views a reference patch with a fixed reflectance and illumination, and a match
 22 patch under illumination that varies from trial to trial. The observer adjusts the
 23 reflectance of the match patch until it appears to be the same as the reflectance of the
 24 reference patch. The match setting r_m can be plotted as a function of the illuminance
 25 i_m at the match patch (Figure 2(b)). In Appendix A we show that for an observer with
 26 a fixed Thouless ratio τ , this plot is a straight line on log-log axes, with slope $\tau - 1$.
 27 Thus such a plot gives a straightforward visual representation of the observer's degree of
 28 lightness constancy. Two useful benchmarks are the line for perfect constancy ($\tau = 1$),
 29 with slope zero, and the line for luminance matching ($\tau = 0$), with slope -1.

30 Previous work

31 Experiments on visual perception often use computer-generated stimuli in order to
 32 maintain precise control over key image features. In fact, some experimental designs
 33 would be difficult to implement at all without computer-generated images, such as those
 34 using inconsistent lighting cues (Ostrovsky, Cavanagh, & Sinha, 2005; Wilder, Adams,
 35 & Murray, 2019) or inconsistent depth cues (Landy, Maloney, Johnston, & Young,

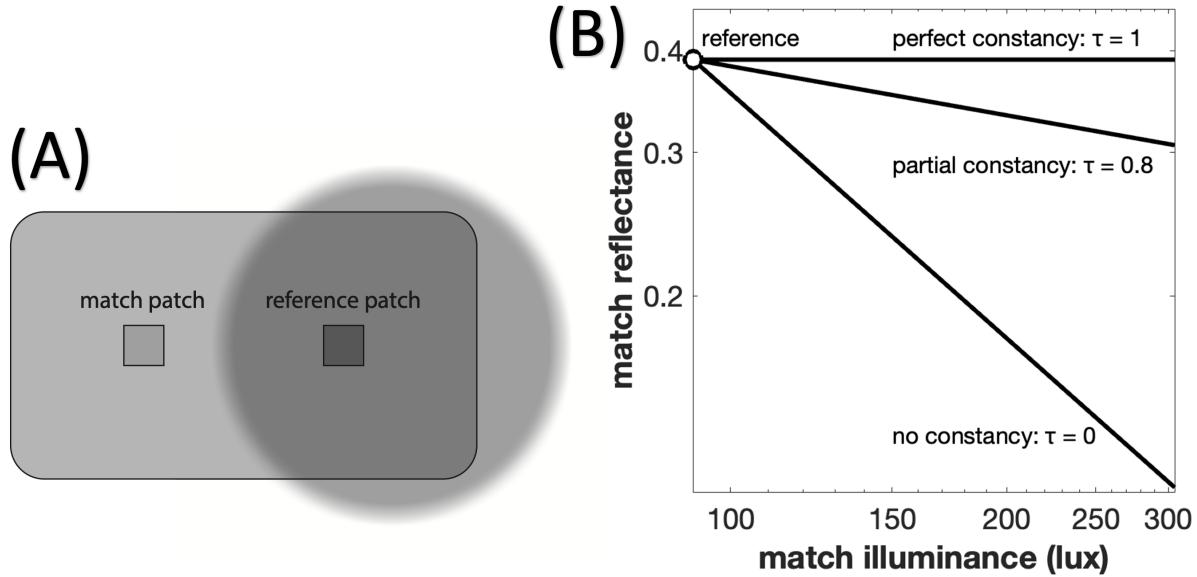


Figure 2. A typical asymmetric lightness matching experiment. (a) The reference patch has a fixed reflectance and illumination level. The match patch has an adjustable reflectance, and is shown under a different illumination level. The observer adjusts the match patch so that it appears to have the same 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.

- ¹ 1995), or using scenes rendered without interreflections (Blok, Kersten, & Hurlbert, 1999). VR extends this control further, and allows the experimenter to create immersive and interactive but well-controlled virtual environments.
- ⁴ 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?
- ⁵ Clearly the answer will depend on the details of rendering and display methods.
- ⁶ Previous work suggests, though, that even when care is taken to render stimuli realistically, perception of real and virtual environments can differ substantially, and
- ⁷ 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.
- ⁸ 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

1 accommodation cues. [Creem-Regehr, Stefanucci, and Bodenheimer \(2022\)](#) review work
2 showing that observers underestimate egocentric distance in virtual environments
3 compared to real environments.

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

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

1 Methods

2 Participants

3 There were twelve participants, all of whom were unaware of the purpose of the
4 experiment. Ten were female, two were male, and ages ranged from 19 to 35 years.
5 Participants gave written informed consent, and were paid for their participation. All
6 reported normal or corrected-to-normal acuity, and reported no known anomalies in
7 colour vision. All procedures were approved by the Office of Research Ethics at York
8 University.

9 Stimuli

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

29 We measured the reflectance of parts of the apparatus using a photometer (model
30 LS-110; Konika Minolta, Tokyo, Japan) and a 99% diffuse reflectance standard
31 (Spectralon SRS-99-020; Labsphere, North Sutton, NH). We measured the luminance L
32 of a target surface location, and the luminance L_{99} of the reflectance standard placed at
33 the same location. Both measurements were made under the illuminant used in the
34 experiment (see below). We calculated the reflectance of the target location as
35 $r = 0.99 L / L_{99}$.

36 The apparatus and room were illuminated by overhead LED lights and a data
37 projector. The LED lights were seven computer-controlled smart bulbs (Philips Hue,

1 White and Color Ambiance E26, 1100 lumen; Signify N.V., Eindhoven, Netherlands) in
 2 an enclosed, spherical cloth lampshade of diameter 55 cm, located above, behind, and to
 3 the left of the observer (Figure 3). The colour and brightness of the LED lights were
 4 adjusted so that the 99% reflectance standard, placed at the centre of the panel
 5 apparatus, had an *xy* chromaticity of (0.30, 0.32) and a luminance of 37 cd/m² as
 6 measured by a spectrophotometer (SpectraScan PR-655; JADAK, North Syracuse, NY).

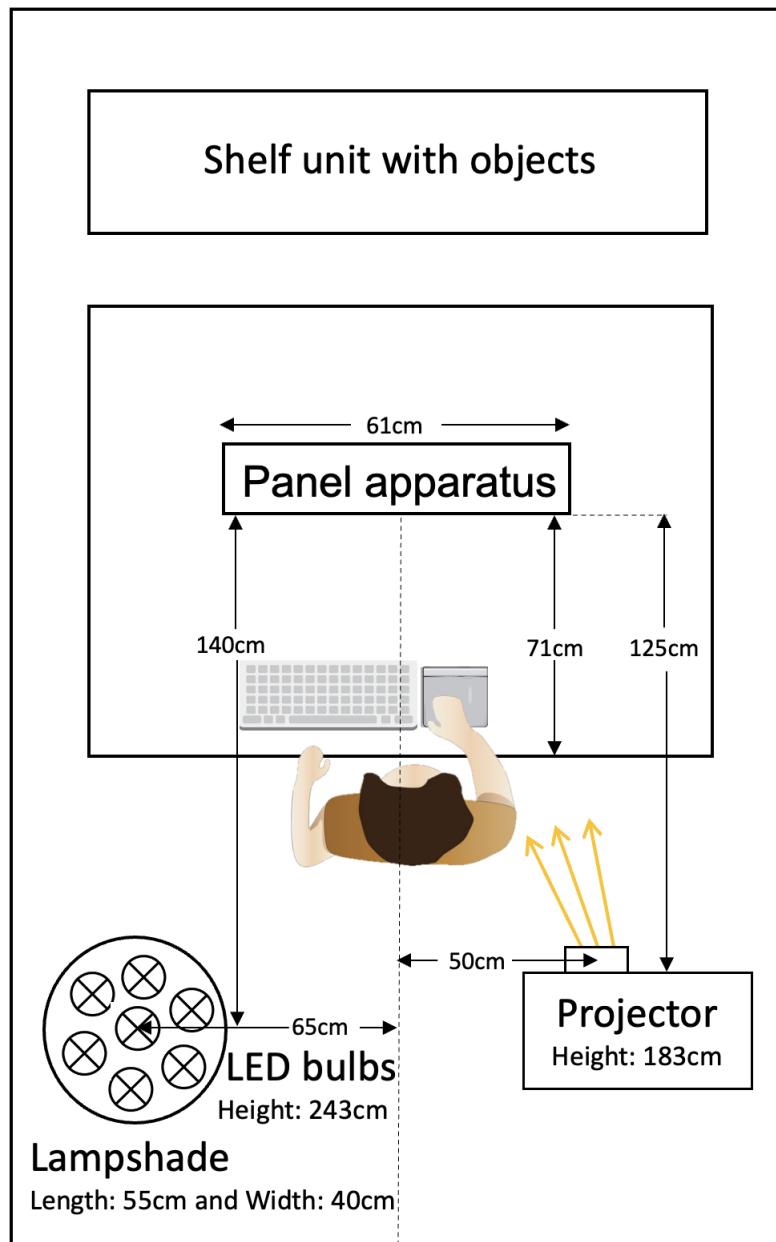


Figure 3. A schematic top-down view of the testing room and apparatus in the physical environment.

7 The data projector was located above, behind, and to the right of the observer,
 8 and directed towards the panel apparatus on the table (model CP-EX252N; Hitachi
 9 Ltd, Tokyo, Japan; Figure 3). We used the photometer and reflectance standard to

1 measure the mapping from the greyscale RGB value displayed on the projector to the
 2 illuminance at the two apertures in the panel apparatus. The *xy* chromaticity of the
 3 projected light at the centre of the panel apparatus was (0.30, 0.32). We displayed a
 4 two-tone image on the data projector in order to create different uniform illumination
 5 intensities on the left and right halves of the panel apparatus (Figure 4(a)). (Specific
 6 illuminances are given below under Procedure.) The resulting illumination boundary
 7 extended beyond the panel apparatus, and created a shadow-like boundary on the table
 8 and surrounding furniture as well. We defocused the projector so that no pixelation was
 9 visible in the projected image, and to blur the lighting boundary, which as a result was
 10 penumbra-like and unlikely to be mistaken for a reflectance boundary.

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

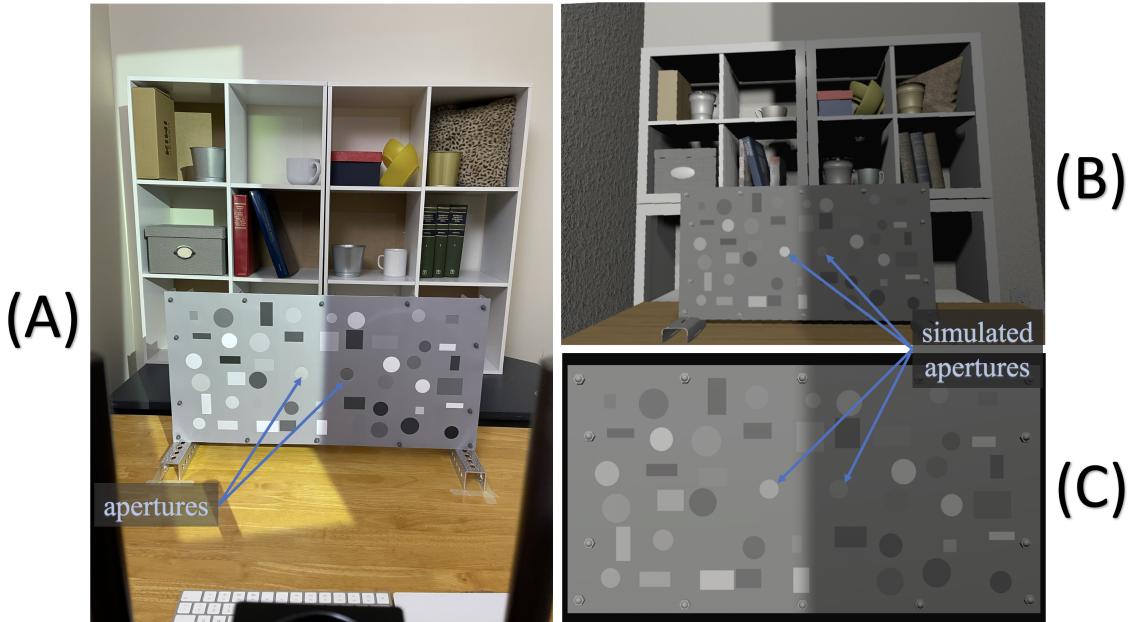


Figure 4. (a) Photograph of the physical room and apparatus, taken from the viewing position used by observers in the experiment. (b) A screen capture of the VR environment. (c) The panel apparatus rendered and displayed on a flat-panel monitor.

14 VR environment.

15 The observer viewed a scene in an Oculus Rift S headset (88° FOV, 1280×1440
 16 pixel resolution per eye), driven by an NVIDIA GeForce GTX 1060 graphics card on a
 17 PC running Windows 10. The scene (Figure 4(b)) was modelled after the room and
 18 apparatus in the physical environment, and rendered in Unity using the Built-in Render
 19 Pipeline (Unity Technologies, 2020, Version 2019.3.3f1). The virtual panel apparatus
 20 was carefully matched to the physical apparatus: the simulated size, position, and
 21 reflectance pattern of the virtual panel were the same as in the physical apparatus, as
 22 were the simulated size and positions of the two apertures. The panel apparatus was

1 rendered as a Lambertian material (Unity material type Legacy/Diffuse). The rest of
2 the virtual scene was also matched to the physical scene, though not as rigorously. The
3 virtual room contained walls, furniture, and objects that we assigned approximately the
4 same simulated size, position, chromaticity, and material properties as their physical
5 counterparts. Size and position were replicated by assigning the virtual objects
6 simulated sizes that matched the measured sizes of the physical objects. Chromaticity
7 and material properties were replicated by adjusting the virtual material properties
8 until the virtual objects appeared approximately the same as the physical objects. We
9 also measured luminance at a coarse grid of locations and adjusted the virtual scene so
10 that the luminances of the rendered RGB values were approximately proportional.

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

25 We calibrated the VR headset to ensure that the luminances displayed were
26 proportional to rendered achromatic RGB values (i.e., the three integers in the range
27 0–255 that represent each pixel). Unity’s Post-Processing Stack includes a feature called
28 the ‘LUT Texture’ that passes rendered RGB values through a nonlinearity specified by
29 the user. We used the LS-110 photometer to measure the nonlinear mapping from
30 achromatic RGB value to physical luminance without the LUT Texture mechanism. We
31 then used the LUT Texture to pass rendered achromatic RGB values through a
32 compensating nonlinearity, with the result that physical luminance was proportional to
33 the rendered achromatic RGB values. We describe this calibration procedure more
34 thoroughly in a separate publication ([Murray, Patel, & Wiedenmann, 2022](#)).

35 **Flat-panel environment.** Observers viewed the same virtual panel apparatus as
36 in the VR environment, but displayed on a flat-panel LCD monitor (Dell UltraSharp
37 27" U2719D; Dell, Round Rock, TX; Figure 4(c)) in a dark room instead of a VR
38 headset. The physical monitor was placed at a viewing distance of 71 cm, where its
39 active area (59.5 cm × 43.5 cm) subtended the same horizontal visual angle as the

1 panels in the physical and VR environments. The stimulus was rendered by the same
2 Unity program used in the VR environment, with a virtual camera positioned so that
3 the panel apparatus filled the full width of the monitor. We measured the monitor's
4 gamma function using the LS-110 photometer, and linearized the mapping from
5 achromatic RGB to luminance using the same post-processing method as in the VR
6 environment. The luminances of the various regions of the panel were proportional to
7 the corresponding luminances in the physical and VR conditions.

8 Procedure

9 Before starting the experiment, the observer made a few practice lightness matches
10 in the physical condition (see below for details). This familiarized them with the task,
11 and providing instructions while they used the physical apparatus made it easier to
12 convey that the goal was to match the perceived shade of grey paper visible through the
13 apertures (i.e., lightness), and not some other stimulus property such as illumination or
14 luminance. Performance in matching tasks can be highly dependent on instructions
15 ([Blakeslee et al., 2008](#)), so for consistency the experimenter read the instructions from a
16 prepared script. After the instructions and practice trials, the observer ran in the
17 physical, VR, and flat-panel conditions in a randomly chosen order. Each observer
18 completed the experiment on a single day, with breaks of at least 20 minutes between
19 the three conditions.

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

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

1 the spacebar on a keyboard to indicate that they had completed the match. A short
2 beep acknowledged their response, and then the next trial began.

3 There were three reference reflectances and five match illuminances (given above).
4 Each reflectance-illuminance combination was repeated five times with the left aperture
5 as the reference, and five times with the right as the reference, for a total of
6 $3 \times 5 \times (5 + 5) = 150$ trials.

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

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

31 **Flat-panel condition.** This condition was largely the same as the physical
32 condition, except that the observer viewed a rendering of the panel apparatus on a
33 flat-panel LCD monitor. Only the panel apparatus, and no other element of the scene,
34 was shown on the monitor, as the goal of this condition was to imitate a typical
35 psychophysical experiment where only the primary stimulus elements are displayed,
36 e.g., Allred, Radonjić, Gilchrist, and Brainard (2012); Anderson and Winawer (2005);
37 Toscani, Zdravković, and Gegenfurtner (2016). The monitor was placed on a table in a
38 dark and otherwise empty testing room. The bookcase and objects that were behind
39 the testing apparatus in the physical condition were not present in this condition. Head

1 position was stabilized by a chinrest. The simulated reference reflectances were the
 2 same as in the other conditions. The simulated illuminances were higher than but
 3 proportional to those in the other two conditions: the simulated reference illuminance
 4 was 286 lx, and simulated match illuminances were 358, 447, 559, 699 and 874 lx. The
 5 observer adjusted the match reflectance using two joysticks on a gamepad, in a manner
 6 similar to the VR condition, and pressed a button on the gamepad to indicate that they
 7 had completed the match.

8 Results

9 Figure 5 shows two typical observers' match settings as a function of illuminance
 10 at the match patch. (Results for the remaining observers are shown in Appendix B.)
 11 The straight lines are sum-of-squares fits, constrained to pass through the points
 12 representing the reference stimuli, which are indicated by open circles. As explained
 13 above, the observer's Thouless ratio is given by one plus the slope of the fitted line, and
 14 this Thouless ratio is shown at the right of each line. In all three environment
 15 conditions, observers' match settings were well fit by straight lines, and match
 16 reflectance was only moderately affected by illuminance. (Code that produces this
 17 figure, as well as all other figures and analyses reported in the Results and Discussion
 18 sections, is provided as Supporting Information at
 19 <https://doi.org/10.17605/OSF.IO/7EUYZ.>)

20 Figure 6(a) shows Thouless ratios for all twelve observers in all conditions. Each
 21 set of three small blue dots connected by straight lines represents an individual
 22 observer, and the larger red dots represent mean Thouless ratios across observers. The
 23 mean Thouless ratio was 0.87 in the physical condition, 0.83 in the VR condition, and
 24 0.79 in the flat-panel condition. Table 1 reports mean Thouless ratios in all conditions,
 25 as well as bootstrapped standard errors and 95% confidence intervals.

Table 1

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

		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

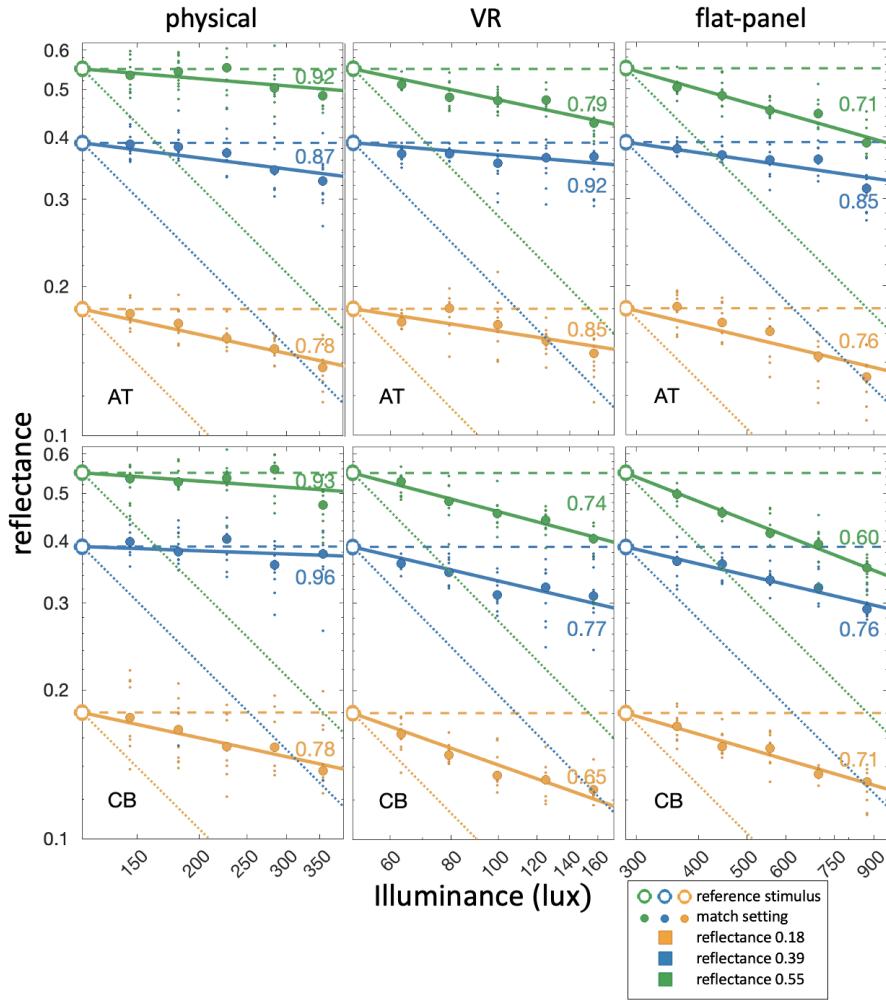


Figure 5. Results for two typical observers in the lightness matching task. Small dots represent reflectance matches on individual trials, and large filled dots represent the median match across trials. Straight lines are sum-of-squares fits, constrained to pass through the open circles that represent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

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

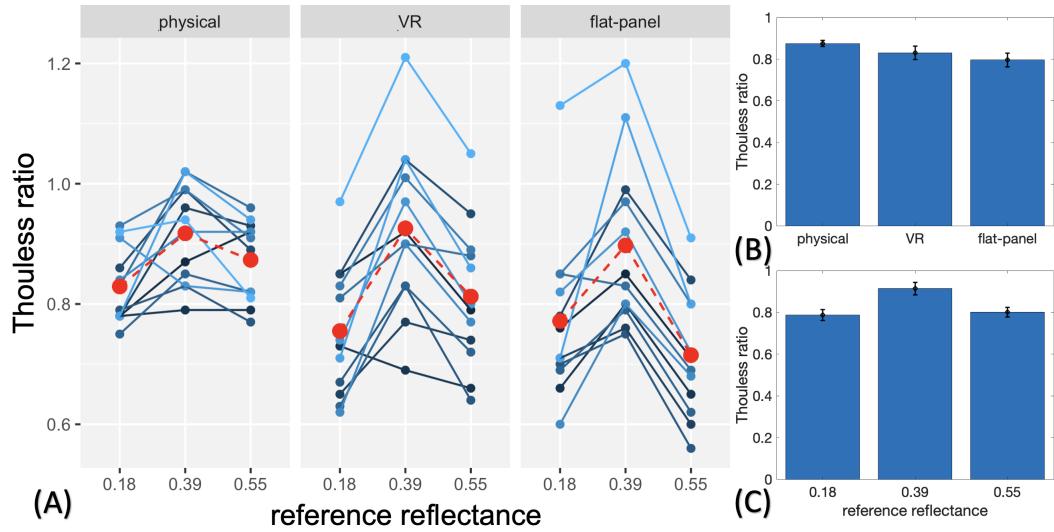


Figure 6. (a) Thouless ratios as a function of environment and reference reflectance. Each triad of blue dots connected by straight lines represents a single observer, and red dots represent means across observers. (b) Mean Thouless ratios by environment. (c) Mean Thouless ratios by reference reflectance.

1 display types. To investigate this hypothesis, we plotted Thouless ratios in the
 2 flat-panel condition against the corresponding Thouless ratios in the VR condition, for
 3 each participant and reference reflectance (Figure 7). There was a clear linear
 4 relationship between Thouless ratios for the two display conditions. A correlation test
 5 using Bonferroni corrected significance level of $p = 0.05/3 = 0.017$ showed significant
 6 Pearson correlations between Thouless ratios in the VR and flat-panel conditions for all
 7 three reference reflectances: reference reflectance 0.18 ($r(34) = 0.85, p < .017, 95\% \text{ CI}$
 8 $[0.54, 0.96]$), reference reflectance 0.39 ($r(34) = 0.91, p < .017, 95\% \text{ CI} [0.71, 0.98]$), and
 9 reference reflectance 0.55 ($r(34) = 0.93, p < .017, 95\% \text{ CI} [0.76, 0.98]$).

10 As the Levene's test indicated non-homogeneous variances, we compared mean
 11 Thouless ratios across conditions using bootstrap tests in order to avoid making strong
 12 distributional assumptions. We created 10^6 bootstrapped replications of the
 13 experiment, where each replication was based on twelve observers randomly selected
 14 with replacement from the twelve observers in the experiment. For each bootstrapped
 15 replication we calculated the mean Thouless ratio across observers for each of the three
 16 environments (physical, VR, flat-panel) and three reference reflectances (0.18, 0.39,
 17 0.55). (See Appendix C for further details of the bootstrapped significance tests.) A
 18 two-tailed comparison of the bootstrapped means showed no significant difference
 19 between the physical and VR conditions ($p = 0.12$). However, the physical condition
 20 had a significantly higher mean than the flat-panel condition ($p < 0.05$), and similarly,
 21 the VR condition had a significantly higher mean than the flat-panel condition ($p <$
 22 0.001). Furthermore, the mean for reference reflectance 0.39 was significantly greater
 23 than the means for reference reflectances 0.18 and 0.55 ($p < 0.001$ in both cases), but

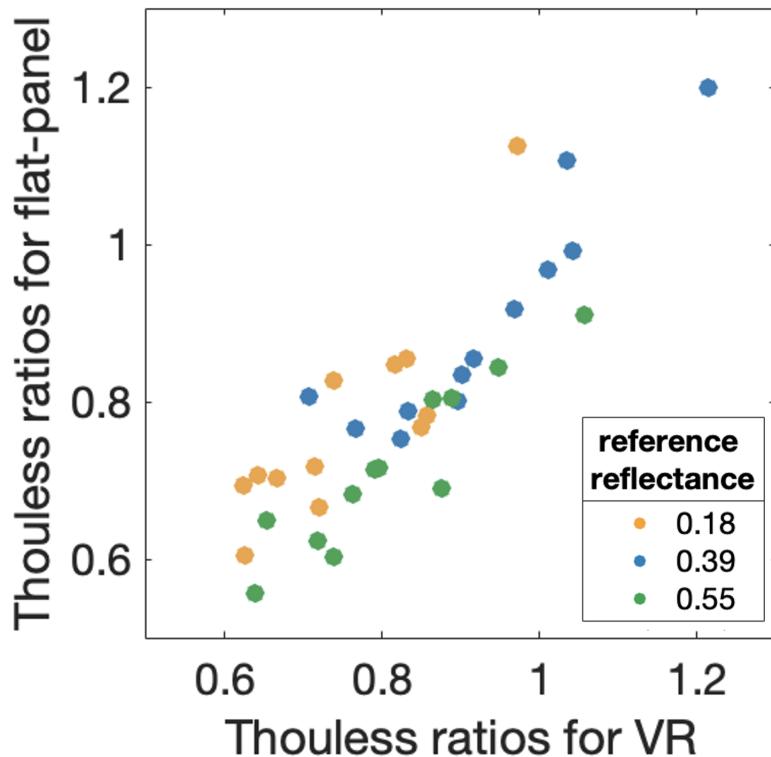


Figure 7. Thouless ratios in the VR and flat-panel conditions are strongly correlated.

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]).

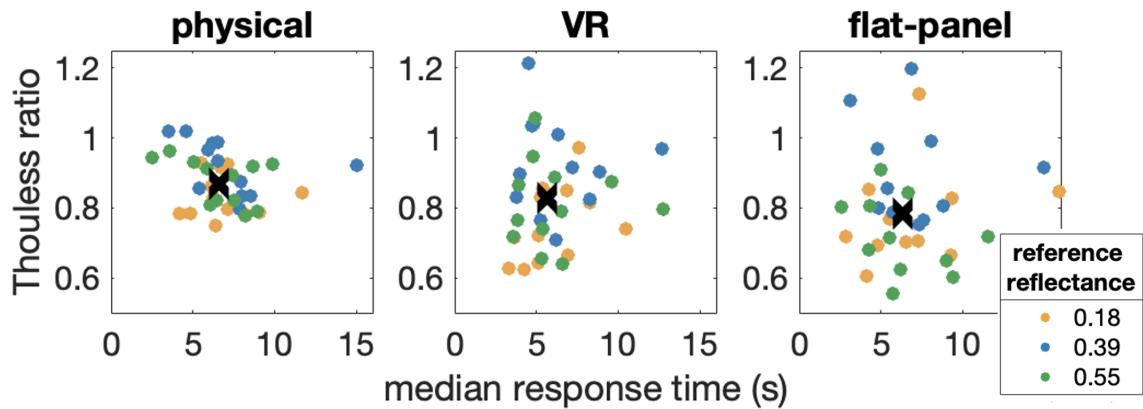


Figure 8. Thouless ratio versus median response time. The black 'x' indicates the median response time and Thouless ratio for each environment. The standard error for the medians was approximately the same size as the marker itself, and so are not displayed in this plot.

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 colour 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 colour 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, Guarnera, Ghosh, Denk, & Glencross, 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, Graf, & Ernst, 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) (Figure 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 (4) 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

¹ constraint on observers' match settings.

² We explored this hypothesis further by developing a simple model where an
³ observer's match setting was determined by both an illumination discounting stage and
⁴ a penalty on match settings that produced opposite contrast polarities at the reference
⁵ and match patches. We did not numerically optimize the fit of this model to our data;
⁶ rather, our goal was to show qualitatively that the contrast-polarity heuristic can
⁷ produce inflated Thouless ratios for low-positive-contrast stimuli, as observed in our
⁸ experiment with human observers. The Supporting Information includes a MATLAB
⁹ implementation of the model.

¹⁰ The first component of the model used equation (8) to calculate an initial match
¹¹ reflectance r_{m*} that was not subject to the penalty on opposite-polarity matches
¹² described below. This calculation assumed a Thouless ratio $\tau = 0.8$, which was the
¹³ average value for human observers for reference reflectances 0.18 and 0.55. τ was
¹⁴ therefore a free parameter of the model. To make this step more realistic as a model of
¹⁵ human behaviour, the model added a random perturbation to the estimate. The
¹⁶ random component was a sample from a normal distribution with mean zero and
¹⁷ standard deviation $\sigma_1 = 0.03$, which was the median standard deviation of human
¹⁸ observers' match settings.

¹⁹ The second component of the model chose a match reflectance r_m by minimizing
²⁰ an objective function with two terms. The first term f_1 was the squared difference
²¹ between r_m and the match reflectance r_{m*} described in the previous paragraph:

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

²² The second term f_2 penalized match settings r_m that gave the match patch a different
²³ local contrast polarity than the reference patch. That is, on trials where the reference
²⁴ reflectance r_{ref} was higher than the background reflectance r_{bg} , the model penalized
²⁵ match reflectance settings r_m that were lower than the background reflectance r_{bg} , and
²⁶ vice versa. The penalty term was

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

²⁷ Here $F(x, \alpha, \beta)$ is the Weibull cumulative distribution function, and $\text{sgn}(x)$ is the
²⁸ signum function:

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

²⁹ We plot the penalty function f_2 in Figure 9(a). α is a scale parameter that determines
³⁰ the match reflectance at which the penalty function saturates; we used $\alpha = 0.05$. β is a

- 1 shape parameter; we used $\beta = 2$. $F(x, \alpha, \beta)$ is nonzero only when $x > 0$, so the sgn
 2 factor causes the penalty function to penalize positive values of $r_m - r_{bg}$ when
 3 $r_{ref} < r_{bg}$, and to penalize negative values when $r_{ref} > r_{bg}$.

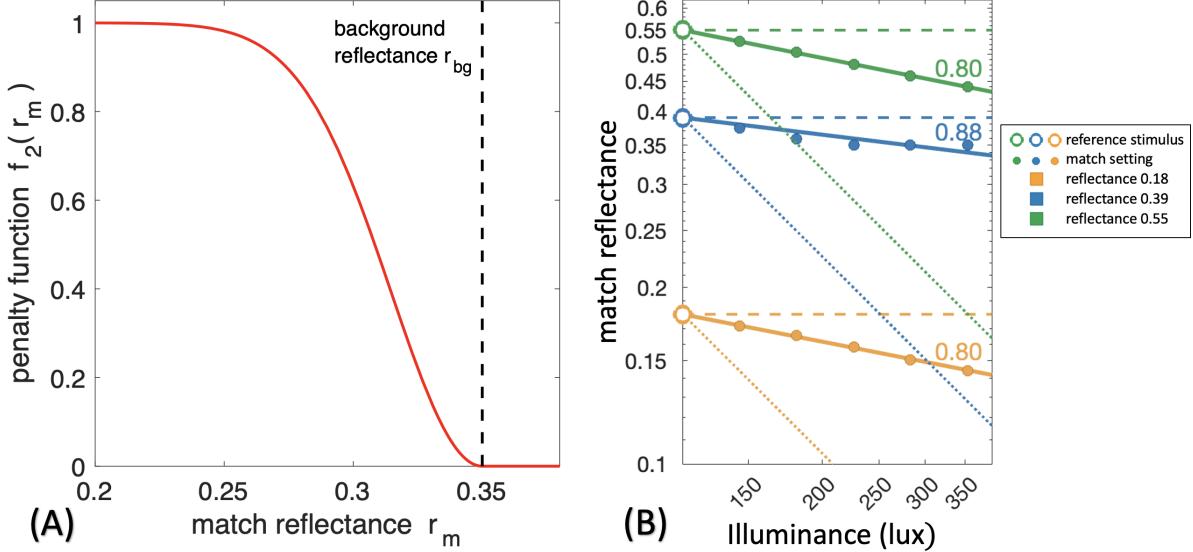


Figure 9. (a) The penalty function f_2 . (b) The model's median match settings and the corresponding Thouless ratios.

- 4 The complete objective function was

$$f(r_m) = f_1(r_m) + w f_2(r_m) \quad (7)$$

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

9 We computed 1000 match reflectance settings (simulating 1000 trials) for each
 10 combination of reference reflectance and match illuminance used in the experiment with
 11 human observers. Figure 9(b) shows the results. The model had a Thouless ratio of 0.8
 12 for reference reflectances 0.18 and 0.55; the penalty term had little effect in these cases,
 13 where the reference reflectances were far above or below the background reflectance of
 14 0.35. The model had a Thouless ratio of 0.88 for the reference reflectance 0.39; here the
 15 penalty term favoured positive-contrast match settings, resulting in a higher Thouless
 16 ratio. Note that the penalty term was not a hard constraint: even with reference
 17 reflectance 0.39, the model sometimes produced match settings lower than the
 18 background reflectance (0.35), as was the case with human observers as well. These
 19 Thouless ratios are practically identical to those reported above for human observers,
 20 and overall these results show that the contrast-polarity heuristic is a plausible
 21 explanation of the higher Thouless ratios that we observed for the reference reflectance

¹ (0.39) that was close to the background reflectance.

² The contrast-polarity model we have presented here is somewhat ad hoc, as it
³ simply imposes a penalty on opposite-polarity matches. A more thorough and
⁴ satisfactory account would motivate this penalty term, for example in terms of natural
⁵ scene statistics. However, here our goal has been to go beyond speculation regarding
⁶ why some reflectance matches were markedly higher than others and to show that, in
⁷ principle, some form of contrast-polarity penalty can quantitatively account for this
⁸ effect. The exact nature of this penalty and the conditions under which it is imposed
⁹ are questions for further study.

¹⁰ *Comparison to previous studies.* Blakeslee et al. (2008) also examined lightness
¹¹ judgements in real and VR environments. The goal of their study, however, was to
¹² examine a range of simultaneous contrast effects, not to measure lightness constancy, or
¹³ to systematically compare performance in real and VR environments. Accordingly, their
¹⁴ stimuli were simple, classic simultaneous contrast figures, and did not provide rich
¹⁵ lighting cues. As a result, their observers reported that lightness judgements were
¹⁶ effortful, conscious calculations, rather than spontaneous judgements of perceived
¹⁷ surface reflectance. For these reasons, Blakeslee et al.'s results are not immediately
¹⁸ comparable to ours, but it is still worth noting that their observers made similar
¹⁹ lightness matches in real and VR environments, consistent with our findings.

²⁰ How consistent is our main finding, namely that there was no significant difference
²¹ in lightness constancy between real and VR environments, with the previous studies
²² reviewed in the introduction? There may seem to be a discrepancy between our findings
²³ and Morgenstern et al. (2014), which found substantial differences between real and
²⁴ virtual environments. However, there are several possible reasons for these divergent
²⁵ findings. First, as noted earlier, Morgenstern et al. compared lightness constancy
²⁶ between two previous studies that were designed and run independently, and had many
²⁷ stimulus differences (Bloj et al., 2004; Boyaci et al., 2003). As a result, Morgenstern et
²⁸ al.'s comparison is not as reliable as a comparison between experiments with similar
²⁹ stimuli and tasks, as in the present study.

³⁰ Second, Boyaci et al. (2003), which provided data for the virtual part of
³¹ Morgenstern et al.'s comparison, used a stationary, custom-built stereoscopic display,
³² not a VR headset. It may be that the immersive VR environment in the present study
³³ contributed to good lightness constancy.

³⁴ Third, and we believe most importantly, the task in our study was fundamentally
³⁵ different from the one in Morgenstern et al. In our study, the task was to match the
³⁶ reflectance of two patches on a single frontoparallel surface, surrounded by the same
³⁷ background reflectance, and separated by a single shadow boundary. In the studies
³⁸ examined by Morgenstern et al., the task was to view an isolated reference patch in a
³⁹ complex scene, rotated to a new 3D orientation on each trial, and to choose a lightness

1 match from a separate frontoparallel palette of match patches in a different part of the
2 scene. It seems plausible that the latter task was intrinsically more demanding, as it
3 required observers to compensate for the intensity of incident light at the 3D orientation
4 of the reference patch (which changed from trial to trial), and to compare the resulting
5 reflectance estimate to a palette at a different orientation and location. It may be that
6 observers achieve good lightness constancy in VR environments on simple tasks like
7 ours, but that more demanding tasks like those examined by Morgenstern et al. reveal
8 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

17 The data and analysis code for all experiments are available at
18 <https://doi.org/10.17605/OSF.IO/7EUYZ>.

Appendix A

1 We assume that all surfaces are Lambertian. Let the reference stimulus have reflectance
 2 r_1 and illuminance i_1 , so following equation (1) for a Lambertian surface, the luminance
 3 is $\ell_1 = r_1 i_1 / \pi$. Let the observer's match setting at the test stimulus be reflectance r_m ,
 4 under illuminance i_m , and so have luminance $\ell_m = r_m i_m / \pi$. If the observer has no
 5 lightness constancy and simply matches the luminance of the reference and test stimuli,
 6 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
 7 expression for r_0 into equation (3) and solve for the match setting $\log r_m$, we find

$$\log r_m = (\tau - 1)(\log i_m - \log i_1) + \log r_1 \quad (8)$$

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

Appendix B

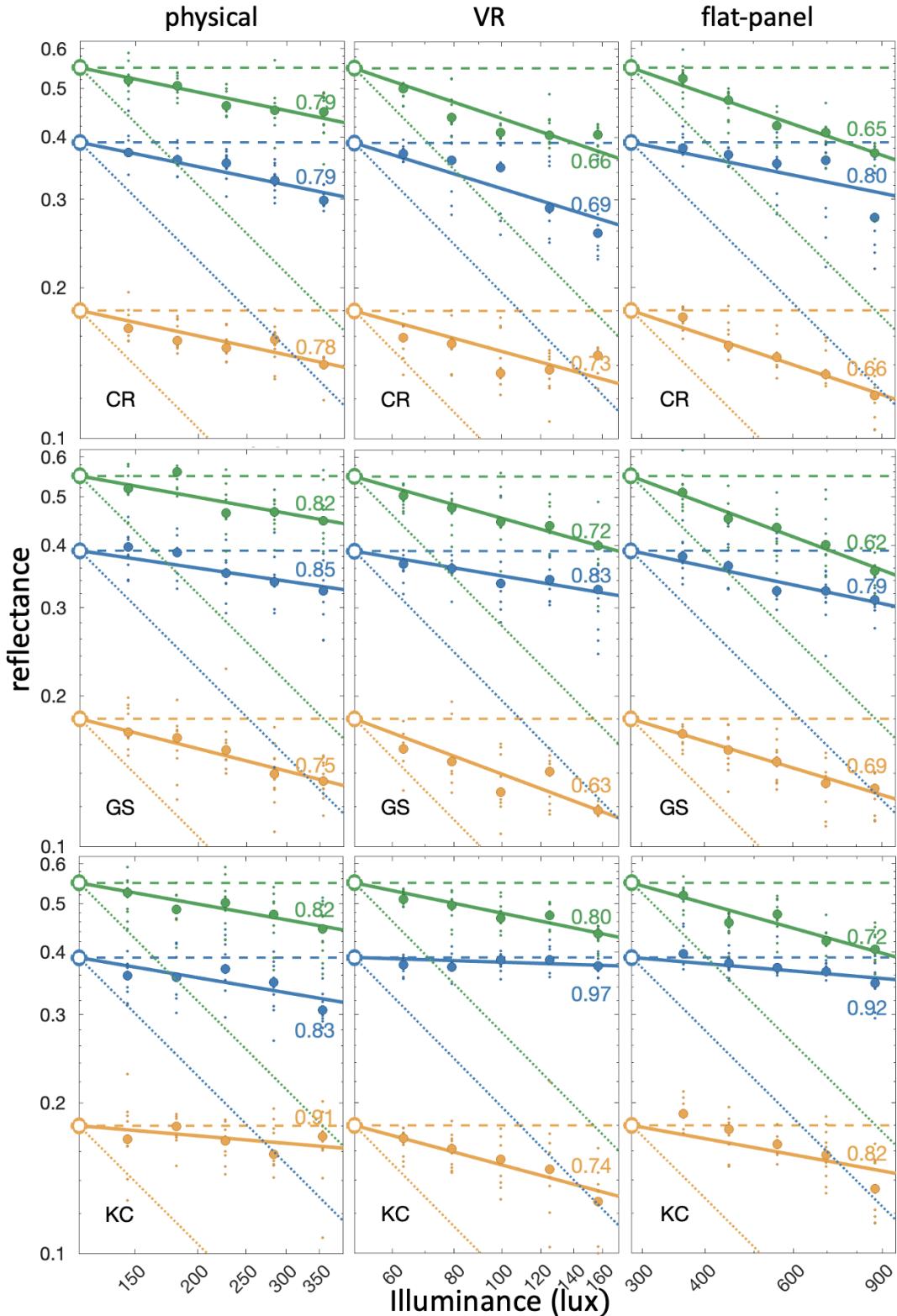


Figure B1. Results for additional observers in the lightness matching task. See caption of Figure 5 for details.

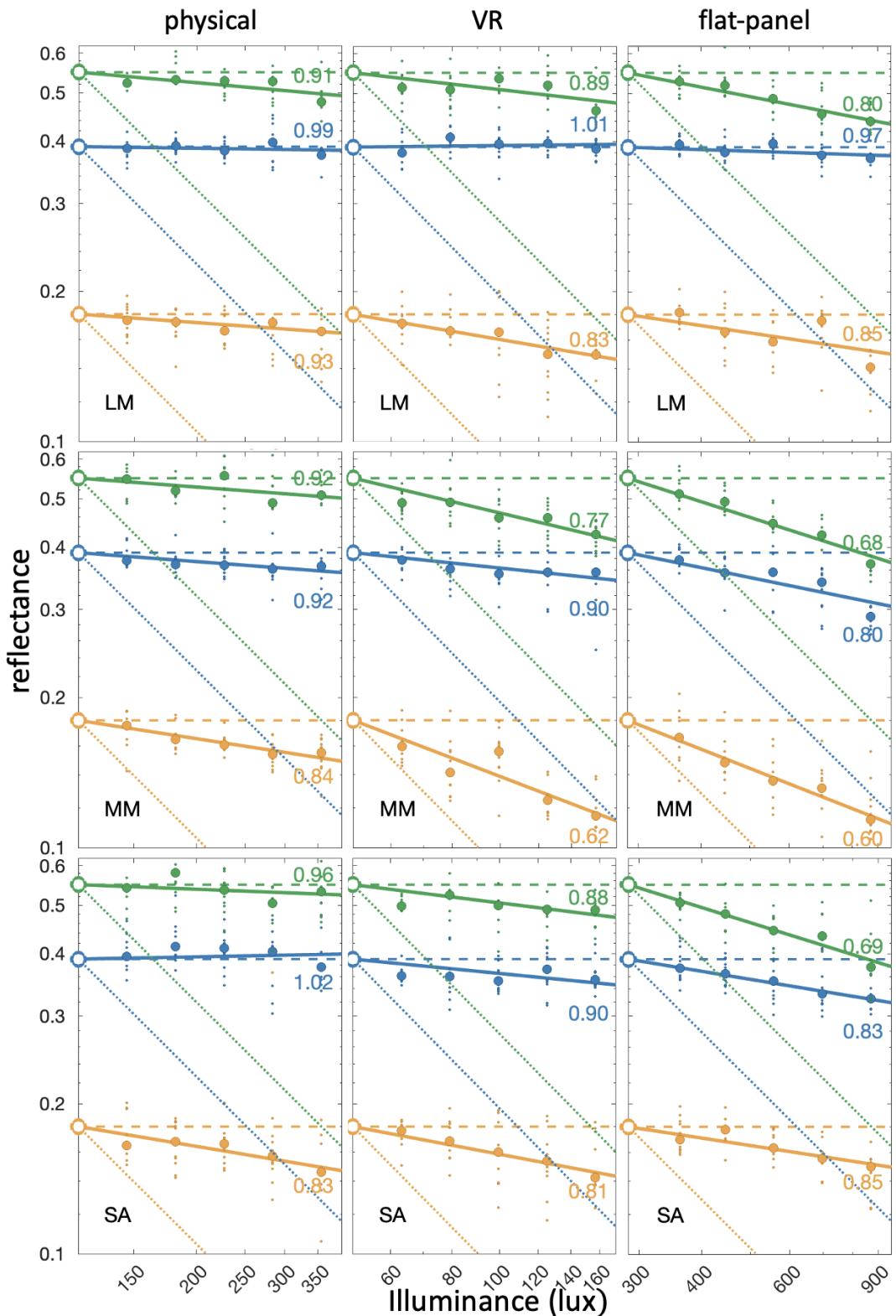


Figure B2. Results for additional observers in the lightness matching task. See caption of Figure 5 for details.

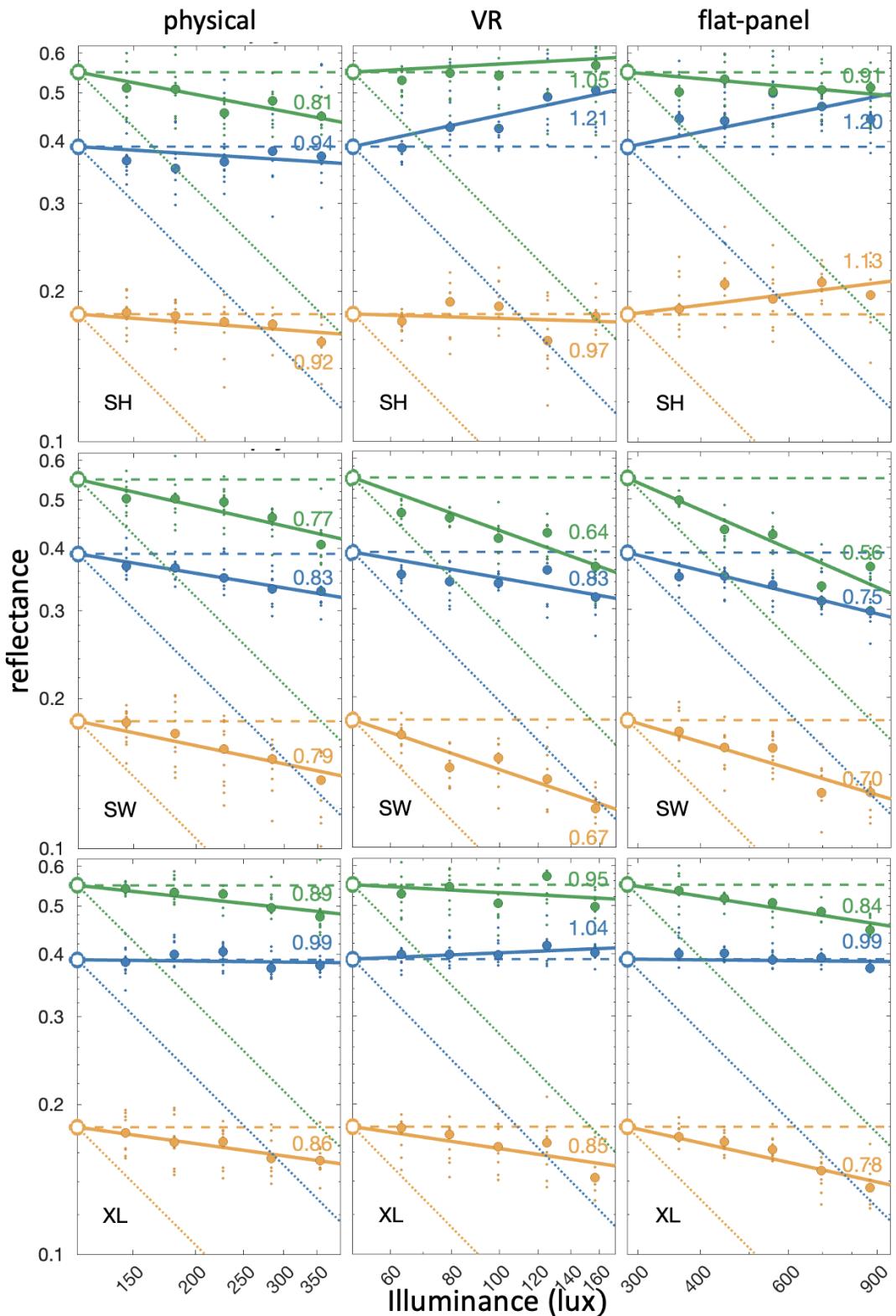


Figure B3. Results for additional observers in the lightness matching task. See caption of Figure 5 for details.

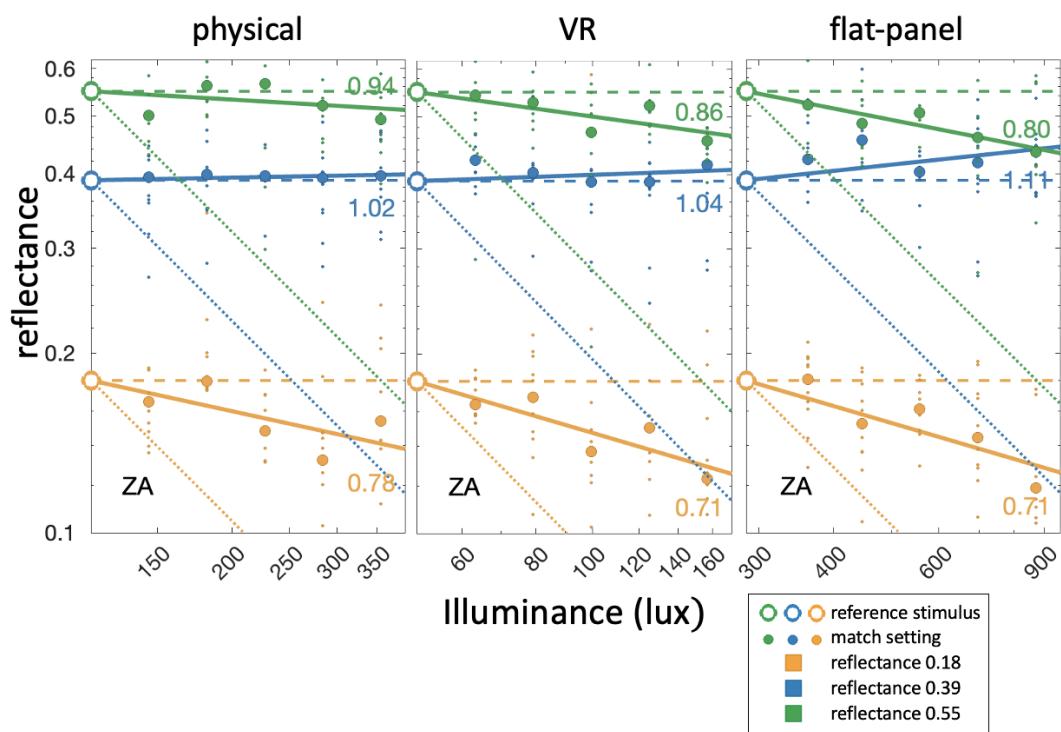


Figure B4. Results for an additional observer in the lightness matching task. See caption of Figure 5 for details.

Appendix C

¹ Here we provide details of the bootstrapped significance tests of mean Thouless ratios.

² Figure 6 shows Thouless ratios for each display method, reference reflectance, and
³ observer. This data can be represented as a $3 \times 3 \times 12$ matrix T_{ijk} , where each entry is
⁴ the Thouless ratio for display method i (an integer from 1 to 3), reference reflectance j
⁵ (also an integer from 1 to 3), and observer k (an integer from 1 to 12). The red dots in
⁶ Figure 6 show mean Thouless ratios across observers, which can be represented as a
⁷ 3×3 matrix M_{ij} , where each element is the average of T_{ijk} for k from 1 to 12.

⁸ On each bootstrap iteration, we simulate a repetition of the experiment by
⁹ creating a new $3 \times 3 \times 12$ matrix $T_{ijk}^{(b)}$ of Thouless ratios. We generate this matrix by
¹⁰ choosing twelve observers with replacement from the twelve observers in the
¹¹ experiment. That is, each 3×3 slice of $T_{ijk}^{(b)}$ for a given value of k is a randomly chosen
¹² 3×3 slice of the original data T_{ijk} , representing the Thouless ratios of a single observer.
¹³ (“With replacement” means that the an observer’s data may appear more than once in
¹⁴ the resampled matrix $T_{ijk}^{(b)}$.) We then calculate the bootstrapped 3×3 matrix of means
¹⁵ $M_{ij}^{(b)}$ by taking the average of $T_{ijk}^{(b)}$ over k from 1 to 12.

¹⁶ We repeat this sampling procedure $B = 10^6$ times, producing B simulated
¹⁷ repetitions of the experiment, represented as $T_{ijk}^{(b)}$ and $M_{ij}^{(b)}$, where b ranges from 1 to B .

¹⁸ This resampled data forms the basis of the bootstrapped significance tests. For
¹⁹ example, to test whether the mean Thouless ratio in condition $i = 1, j = 1$ (say, the
²⁰ physical condition and reference reflectance 0.18) is significantly greater than the
²¹ Thouless ratio in condition $i = 1, j = 2$ (say, the physical condition and reference
²² reflectance 0.39), we find the proportion of bootstrapped samples for which $M_{11}^{(b)}$ is
²³ greater than $M_{12}^{(b)}$. If this is true for at least 95% of the samples, then we conclude that
²⁴ the first mean is significantly greater than the second mean at a significance level of
²⁵ $p < 0.05$.

²⁶ For additional details of bootstrapping methods, see [Efron and Tibshirani \(1994\)](#).

²⁷ We provide MATLAB code that implements these significance tests as Supporting
²⁸ Information.

References

- 1 Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the
2 'light-from-above' prior. *Nature Neuroscience*, 7(10), 1057-1058.
- 3 Adelson, E. H. (2000). Lightness perception and lightness illusions. In M. Gazzaniga
4 (Ed.), *The new cognitive neurosciences* (p. 339-351). The MIT Press.
- 5 Allred, S. R., Radonjić, A., Gilchrist, A. L., & Brainard, D. H. (2012). Lightness
6 perception in high dynamic range images: Local and remote luminance effects.
7 *Journal of Vision*, 12(2), 7-7.
- 8 Anderson, B. L., & Winawer, J. (2005). Image segmentation and lightness perception.
9 *Nature*, 434(7029), 79-83.
- 10 Arend, L. E., Reeves, A., Schirillo, J., & Goldstein, R. (1991). Simultaneous color
11 constancy: papers with diverse munsell values. *Journal of the Optical Society of
12 America A*, 8, 661-672.
- 13 Barron, J. T., & Malik, J. (2015). Shape, illumination, and reflectance from shading.
14 *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8),
15 1670-1687.
- 16 Belhumeur, P. N., Kriegman, D. J., & Yuille, A. L. (1999). The bas-relief ambiguity.
17 *International Journal of Computer Vision*, 35(1), 33-44.
- 18 Berkman, M. I., & Akan, E. (2019). Presence and immersion in virtual reality. In
19 N. Lee (Ed.), *Encyclopedia of computer graphics and games*. Springer.
- 20 Blakeslee, B., Reetz, D., & McCourt, M. (2008). Coming to terms with lightness and
21 brightness: Effects of stimulus configuration and instructions on brightness and
22 lightness judgments. *Journal of Vision*, 8(11), 1-18.
- 23 Bloj, M., Kersten, D., & Hurlbert, A. (1999). Perception of three-dimensional shape
24 influences colour perception through mutual illumination. *Nature*, 402(6764),
25 877-879.
- 26 Bloj, M., Ripamonti, C., Mitha, K., Hauck, R., Greenwald, S., & Brainard, D. H.
27 (2004). An equivalent illuminant model for the effect of surface slant on perceived
28 lightness. *Journal of Vision*, 4(9):6.
- 29 Boyaci, H., Maloney, L. T., & Hersh, S. (2003). The effect of perceived surface
30 orientation on perceived surface albedo in binocularly viewed scenes. *Journal of
31 Vision*, 3(8), 541-553.
- 32 Brainard, D. H. (1998). Color constancy in the nearly natural image. 2. Achromatic
33 loci. *Journal of the Optical Society of America A*, 15, 307-325.
- 34 Brainard, D. H., & Maloney, L. T. (2011). Surface color perception and equivalent
35 illumination models. *Journal of Vision*, 11(5), 1-1.
- 36 Brunswik, E. (1928). Zur Entwicklung der Albedowahrnehmung. *Zeitschrift für
37 Psychologie*, 109, 40-115.
- 38

- 1 Creem-Regehr, S. H., Stefanucci, J. K., & Bodenheimer, B. (2022). Perceiving distance
2 in virtual reality: theoretical insights from contemporary technologies.
3 *Philosophical Transactions of the Royal Society B*, 378(20210456), 1-12.
- 4 Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and
5 Hall/CRC Press.
- 6 Elor, A., & Kurniawan, S. (2020). The ultimate display for physical rehabilitation: a
7 bridging review on immersive virtual reality. *Frontiers in Virtual Reality*,
8 1(585993), 1-17.
- 9 Fairchild, M. D. (2013). *Color appearance models, third edition*. West Sussex, UK: John
10 Wiley & Sons, Ltd.
- 11 Fechner, G. (1860/1966). *Elements of psychophysics*. New York, NY: Holt, Rinehart
12 and Winston.
- 13 Gilchrist, A. (2006). *Seeing black and white*. New York: Oxford University Press.
- 14 Gil Rodríguez, R., Bayer, F., Toscani, M., Guarnera, D., Guarnera, G. C., &
15 Gegenfurtner, K. R. (2022). Colour calibration of a head mounted display for
16 colour vision research using virtual reality. *SN Computer Science*, 3(1), 1–10.
- 17 Greengard, S. (2019). *Virtual reality*. The MIT Press.
- 18 Guarnera, D., Guarnera, G. C., Ghosh, A., Denk, C., & Glencross, M. (2016). Brdf
19 representation and acquisition. *Computer Graphics Forum*, 35, 625-650.
- 20 Hartle, B., & Wilcox, L. M. (2022). Stereoscopic depth constancy for physical objects
21 and their virtual counterparts. *Journal of Vision*, 22(4), 1-19.
- 22 Hibbard, P. B. (2023). Virtual reality for vision science. In *Topics in behavioral
23 neurosciences* (pp. 1–29). Springer.
- 24 Jung, S., & Lindeman, R. W. (2021). Does realism improve presence in VR? Suggesting
25 a model and metric for VR experience evaluation. *Frontiers in Virtual Reality*,
26 2(693327), 1-7.
- 27 Kimura, K., Reichert, J. F., Olson, A., Pouya, O. R., Wang, X., Moussavi, Z., & Kelly,
28 D. M. (2017). Orientation in virtual reality does not fully measure up to the
29 real-world. *Scientific Reports*, 7(1), 1-8.
- 30 Kraft, J. M., & Brainard, D. H. (1999). Mechanisms of color constancy under nearly
31 natural viewing. *Proceedings of the National Academy of Sciences*, 96(1),
32 307–312.
- 33 Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and
34 modeling of depth cue combination: in defense of weak fusion. *Vision Research*,
35 35(3), 389-412.
- 36 Li, Z., Shafiei, M., Ramamoorthi, R., Sunkavalli, K., & Chandraker, M. (2020). Inverse
37 rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf
38 from a single image. *Proceedings of the IEEE/CVF Conference on Computer
39 Vision and Pattern Recognition*, 2475–2484.

- 1 Linowes, J. (2020). *Unity 2020 virtual reality projects: Learn vr development by building*
2 *immersive applications and games with unity 2019.4 and later versions*. Packt
3 Publishing Ltd.
- 4 Marantz, A. (2016). Studio 360. *The New Yorker*, April 25, 2016, pp. 86-94.
- 5 McCluney, R. (1994). *Introduction to radiometry and photometry*. Artech House, Inc.
- 6 Morgenstern, Y., Geisler, W. S., & Murray, R. F. (2014). Human vision is attuned to
7 the diffuseness of natural light. *Journal of Vision*, 14(9), 1-18.
- 8 Murray, R. F. (2021). Lightness perception in complex scenes. *Annual Review of Vision*
9 *Science*, 7, 417–436.
- 10 Murray, R. F., Patel, K. Y., & Wiedenmann, E. S. (2022). Luminance calibration of
11 virtual reality displays in unity. *Journal of Vision*, 22(13):1, 1-9.
- 12 Ong, T., Wilczewski, H., Soni, H., Nisbet, Q., Paige, S. R., Barrera, J., ... Bunnell, B.
13 (2022). The symbiosis of virtual reality exposure therapy and telemental health: a
14 review. *Frontiers in Virtual Reality*, 3(848066), 1-11.
- 15 Ostrovsky, Y., Cavanagh, P., & Sinha, P. (2005). Perceiving illumination inconsistencies
16 in scenes. *Perception*, 34(11), 1301–1314.
- 17 Patel, K., Munasinghe, A., & Murray, R. (2018). Lightness matching and perceptual
18 similarity. *Journal of Vision*, 18(5), 1-18.
- 19 Radonjić, A., Pearce, B., Aston, S., Krieger, A., Dubin, H., Cottaris, N. P., ...
20 Hurlbert, A. C. (2016). Illumination discrimination in real and simulated scenes.
21 *Journal of Vision*, 16(11), 1-18.
- 22 Rzepka, A. M., Hussey, K. J., Maltz, M. V., Babin, K., Wilcox, L. M., & Culham, J. C.
23 (2023). Familiar size affects perception differently in virtual reality and the real
24 world. *Philosophical Transactions of the Royal Society B*, 378(1869), 1-14.
- 25 Sanchez-Vives, M. V., & Slater, M. (2005). From presence to consciousness through
26 virtual reality. *Nature Reviews Neuroscience*, 6(4), 332–339.
- 27 Scarfe, P., & Glennerster, A. (2015). Using high-fidelity virtual reality to study
28 perception in freely moving observers. *Journal of Vision*, 15(9), 3–3.
- 29 Scarfe, P., & Glennerster, A. (2019). The science behind virtual reality displays.
30 *Annual Review of Vision Science*, 5, 529–547.
- 31 Slater, M. (2018). Immersion and the illusion of presence in virtual reality. *British*
32 *Journal of Psychology*, 109(3), 431–433.
- 33 Slater, M., Lotto, B., Arnold, M. M., & Sánchez-Vives, M. V. (2009). How we
34 experience immersive virtual environments: the concept of presence and its
35 measurement. *Anuario de Psicología*, 40, 193–210.
- 36 Thouless, R. (1931). Phenomenal regression to the real object. *British Journal of*
37 *Psychology*, 21(4), 339-359.
- 38 Toscani, M., Zdravković, S., & Gegenfurtner, K. R. (2016). Lightness perception for
39 surfaces moving through different illumination levels. *Journal of Vision*, 16(15),

- 1 21–21.
- 2 Unity Technologies. (2020). *Unity, version 2019.3.3f1*. Retrieved from <https://docs.unity3d.com/2020.1/Documentation/Manual/UnityManual.html>
- 3 Westerman, S. J., & Cribbin, T. (1998). Individual differences in the use of depth cues:
4 implications for computer- and video-based tasks. *Acta Psychologica*, 99, 293-310.
- 5 Wilder, J. D., Adams, W. J., & Murray, R. F. (2019). Shape from shading under
6 inconsistent illumination. *Journal of Vision*, 19(6):2.
- 7 Xie, B., Alghofalli, R., Zhang, Y., Jiang, Y., Lobo, F. D., Li, C., . . . Yu, L.-F. (2021).
8 A review on virtual reality skill training applications. *Frontiers in Virtual Reality*,
9 2(645153), 1-19.
- 10 Xiong, J., Hsiang, E.-L., He, Z., Zhan, T., & Wu, S.-T. (2021). Augmented reality and
11 virtual reality displays: emerging technologies and future perspectives. *Light:
12 Science & Applications*, 10(216), 1-30.
- 13 Zhan, T., Yin, K., Xiong, J., He, Z., & Wu, S.-T. (2020). Augmented reality and virtual
14 reality displays: perspectives and challenges. *iScience*, 23(101397), 1-13.
15