

Lightness perception in augmented reality: Effects of context, background, motion, and disparity

Tasfia Ahsan 

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



Minjung Kim

Independent research scientist, Vancouver, Canada



Richard F. Murray 

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



Lightness constancy—the ability to form a stable percept of surface reflectance despite changes in illumination and context—is critical for the accurate interpretation of visual scenes, including virtual content in optical see-through augmented reality (OST-AR). However, the additive blending of virtual and real-world scenes in AR elevates the luminance of stimuli, which can make it difficult to create a convincing visual impression of low-reflectance surfaces. Here we investigated how a range of visual cues affects lightness perception in OST-AR, with the goal of identifying viewing conditions that expand the range of perceived reflectance for virtual stimuli. In two experiments, participants viewed virtual test patches in a foreground plane and rated their lightness relative to real patches on a paper reference palette. We manipulated four factors: the composition of surrounding elements in the foreground plane, the content and luminance of the background plane, the motion of the background plane, and the relative disparity between foreground and background. We found that the strongest effects on perceived reflectance came from changes to the foreground context: Isolating the test patch or eliminating luminance variation in the surrounding texture substantially reduced the range of perceived reflectance. Surprisingly, even large changes to the background, such as removing texture, introducing motion, or eliminating the depth disparity between foreground and background layers, had much smaller effects. These findings show that lightness perception in OST-AR depends heavily on how the foreground is structured and that including a rich foreground context can help virtual surfaces to generate a wide and realistic range of perceived reflectance.

Introduction

Augmented reality (AR) is an immersive display technology that integrates virtual content with real-world environments. It is increasingly widespread in applications such as medical visualization, industrial training, architectural design, and entertainment (Mekni & Lemieux, 2014). Unlike traditional flat-panel displays, AR allows users to view their physical surroundings with digital elements layered in. This integration can make displayed content appear natural and realistic, but it also creates unique perceptual challenges: To be visually convincing, virtual elements must appear to be embedded in the physical scene and to match its lighting and surface properties (Murdoch, 2025). Even with improvements in rendering and spatial tracking (Itoh, Langlotz, Sutton, & Plopski, 2021), visual inconsistencies remain, and this limits the precision of AR in tasks that require accurate appearance judgments.

A key challenge in AR is creating stimuli that support color constancy, which is the visual system's ability to maintain a stable percept of an object's surface color despite changes in lighting and context (Brainard & Maloney, 2011). An important special case of color constancy is lightness constancy, the ability to perceive achromatic reflectance over changes in lighting and context (Gilchrist, 2006). Achromatic reflectance is the proportion of light in the visible wavelength range that is reflected by a surface (McCluney, 2014). It is often called albedo. Lightness is defined as the perceived achromatic reflectance of a surface, which is seen as black, white, or shades of gray. In optical see-through augmented reality (OST-AR)—a type of AR display where virtual content is additively layered over the real world viewed through a transparent display—the

Citation: Ahsan, T., Kim, M., & Murray, R. F. (2026). Lightness perception in augmented reality: Effects of context, background, motion, and disparity. *Journal of Vision*, 26(9):11, 1–18, <https://doi.org/10.1167/jov.26.9.11>.

<https://doi.org/10.1167/jov.26.9.11>

Received March 5, 2026; published September 28, 2026

ISSN 1534-7362 Copyright 2026 The Authors



luminance of a stimulus is affected by both the content shown on the transparent display and the background scene, making lightness constancy especially fragile. Bleed-through from the background can make it difficult to render dark virtual elements convincingly, particularly in bright environments (Gabbard, Swan, Zedlitz, & Winchester, 2010). Low-reflectance content, such as black text or surfaces, often appears faded or semi-transparent when displayed over bright backgrounds. Thus, the appearance of virtual elements can shift from one context to another, compromising realism and perceptual stability (Itoh et al., 2021). In particular, avatars with dark simulated skin are susceptible to these effects, which can make it difficult to create convincing representations of people in virtual environments (Peck, Good, Erickson, Bynum, & Bruder, 2022).

For these reasons, it would be useful to understand what factors affect lightness perception in AR and, more specifically, what kinds of stimuli generate a wide range of lightness percepts, ranging from black to white. There is a long history of research on lightness perception that can provide starting points for investigating these questions (Gilchrist, 2006; Kingdom, 2011; Murray, 2021). It also seems likely that scission is an important concept for understanding lightness in AR, meaning the ability to perceive a stimulus as consisting of one or more distinct, transparent layers in front of an opaque background. Thus, the large literature on transparency is relevant as well (Metelli, 1970; Singh & Anderson, 2002). In principle, if we can create viewing conditions in AR where the visual system perceptually segments the transparent foreground from the background, the background might have only a small effect on the lightness of foreground elements.

Previous work has demonstrated some important features of lightness perception in AR, relating to the interplay of foreground and background images. For example, Murdoch (2020) examined lightness matches for physical objects when they were viewed through several types of transparent AR overlays, and found that observers were able to partially discount the foreground overlays when judging the background objects, much as they discount lighting and transparency when estimating surface reflectance in more traditional scenes (Adelson, 2000; Gilchrist, 2006). Similarly, He and Thorstenson (2025) showed that the appearance of AR avatars with a range of skin types is modulated by bright backgrounds and, furthermore, that it is not always possible to completely compensate for this effect by adjusting the luminance and color of the avatars. Murdoch (2025) showed that providing foreground context for an avatar, instead of showing the avatar alone, may be an effective way of reducing this influence of bright backgrounds on foreground images. We discuss these studies and additional previous work more thoroughly in the General Discussion.

The present study builds on previous work and explores how several properties of visual scenes affect lightness perception in AR, with the goal of understanding how to generate a wide range of perceived reflectance. We wished to use a stimulus class that was simple enough to control and manipulate precisely but flexible enough to provide a wide range of variations that could potentially affect lightness and scission. We found Voronoi textures to be a good solution to this problem (Figure 1). These textures provide a complex, randomly generated two-dimensional (2D) context for lightness judgments and can vary in their luminance range, as well as the number and size of elements. This allowed us, for example, to vary the “articulation” of stimuli, i.e. the extent to which the stimulus was divided into many regions with different luminances, which is known to have a strong effect on lightness perception (Gilchrist, 2006). Participants judged the lightness of a test patch in a Voronoi texture in the foreground plane of an AR display, superimposed on various kinds of backgrounds. In Experiment 1, we manipulated the complexity and content of the foreground and background, the motion of the background, and depth cues that distinguished between the foreground and background planes. The foreground manipulations were motivated by anchoring theory (Gilchrist, 2006), which holds that lightness is judged relative to the highest luminance in a scene region and predicts that the structure and luminance range of surrounding elements should strongly influence lightness in the foreground plane. The motion manipulation was motivated by evidence that relative motion between surfaces provides a powerful cue for layer separation and scission (Anderson & Winawer, 2005; Anderson & Winawer, 2008). The depth manipulations were motivated by evidence that binocular disparity and depth separation contribute to scission, with surfaces on different depth planes more likely to be perceived as belonging to distinct illumination environments (Gilchrist, 1977; Schirillo, Reeves, & Arend, 1990). In Experiment 2, we introduced additional manipulations of the foreground and background to further explore some results of Experiment 1. To anticipate, we found that the strongest effects on lightness perception came from manipulations of foreground elements surrounding the test patch and that other potential cues for scission between foreground and background, such as relative motion and depth, had surprisingly small effects.

Experiment 1

In Experiment 1, we examined how several properties of visual scenes affect lightness perception in AR, by having participants judge the lightness of a virtual

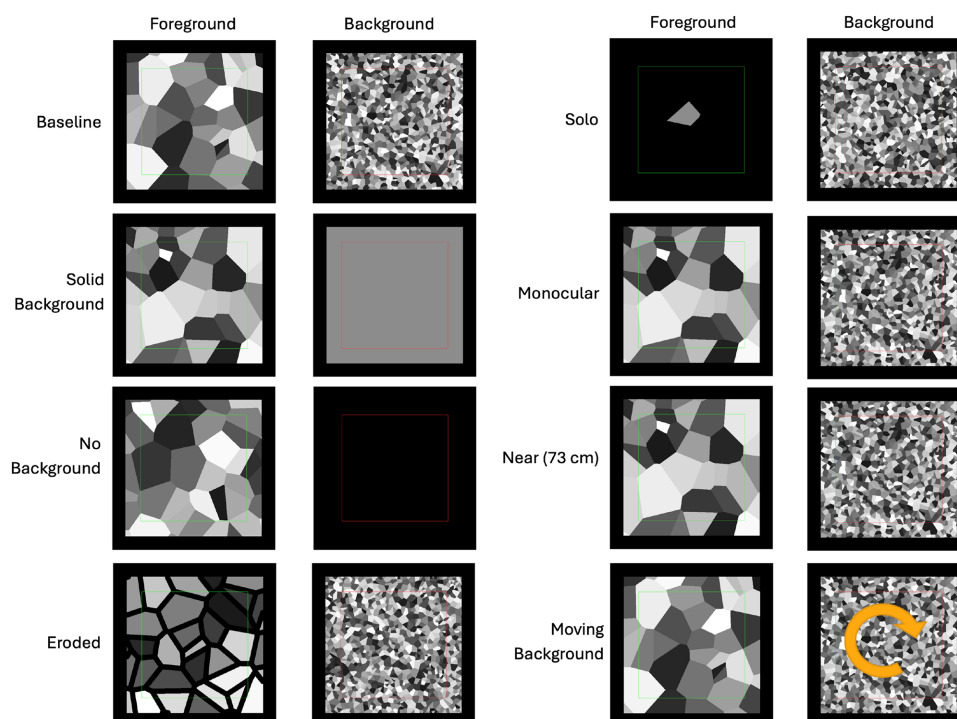


Figure 1. Examples of stimuli in [Experiment 1](#). Each condition is shown as a foreground image (left) and background image (right). Baseline, with Voronoi textures as foreground and background. Solid-Background, with a uniform gray background. No-Background, with a black background. Eroded, with foreground pixels near patch boundaries set to black. Solo, with the foreground showing only the test patch. Monocular, same as Baseline, but participants wore an eye patch over their left eye. Near, same as Baseline, but with the background at a viewing distance of 73 cm rather than 147 cm. Moving-Background, same as Baseline, but with the background moving in a circular path.

test patch in a range of foreground and background contexts. In the Baseline condition ([Figure 1](#)), the test patch appeared in a foreground that was a grayscale Voronoi texture, in front of a background that was a grayscale Voronoi texture with smaller elements. This stimulus created strong segmentation between layers and served as a reference to which we compared results in other conditions.

To examine the role of the background texture, we included a Solid-Background condition, where the background was uniform gray, and a No-Background condition, where the background was uniform black. On an OST-AR display, where virtual foreground imagery is additively combined with the background scene, using a black background effectively makes the stimulus only the foreground image.

To examine the effect of the context surrounding the test patch in the foreground, we included an Eroded condition, where there was a gap between all foreground texture elements, and a Solo condition, where only the test patch appeared in the foreground, with no surrounding texture elements ([Figure 1](#)).

Finally, we included conditions that examined the effect of nonpictorial cues on segmentation between the foreground and background. Motion can be a strong segmentation cue, so we included a

Moving-Background condition, where the background texture moved in a circular trajectory. We also examined the role of binocular depth cues by including a Monocular condition, where one eye was covered by an eye patch, and a Near condition, where the background plane was placed at approximately the same viewing distance as the foreground plane.

Methods

Participants

Twelve naive participants (11 female, 1 male; mean age 22.8 years, range 21–29 years) were recruited from York University, Canada. All reported normal or corrected-to-normal acuity and no known anomalies in depth perception or color vision. Participants provided written informed consent and received an honorarium. Procedures were approved by the York University Human Participants Review Committee.

Apparatus

The experiment was carried out in a room where the only lighting was a lamp that dimly illuminated

the apparatus. We provided this minimal lighting to ensure that participants could comfortably view and select from the Munsell reference patches used to report their lightness judgments, while minimizing uncontrolled luminance variation in the surrounding environment. Under this lighting, the luminance of the whitest regions of the Munsell response palette (see below), located just in front of the participant, was 27.6 cd/m^2 . Head position was stabilized by a chinrest.

Participants viewed stimuli in a custom-built OST-AR display (Figure 2). The apparatus included two 24-inch LCD monitors (Samsung model 24T350, resolution $1,920 \times 1,080$ pixels), one positioned upright and the other facing upward. A beam splitter (Edmund Optics, model 68-268, reflection coefficient 0.75, transmission coefficient 0.25) was placed at an angle of approximately 45° between the two monitors, superimposing their images into a composite AR view. The upward-facing (foreground) monitor was at a fixed optical distance of 73 cm (1.37 diopters) from the viewing position in the chinrest, and the upright (background) monitor was at a distance of either 73 cm (1.37 diopters) or 147 cm (0.68 diopters) in different conditions. This configuration gave the AR display separate foreground and background focal planes, which mitigated vergence-accommodation conflict and was within the visual comfort zone defined by Shibata, Kim, Hoffman, and Banks (2011). The monitors were driven by a PC running Windows 11, MATLAB R2024b (The MathWorks, Inc., Natick, MA, USA),

and the Psychophysics Toolbox (Kleiner, Brainard, & Pelli, 2007).

We made color characterization measurements of the apparatus from the participant's viewing position, using a spectrophotometer (Photo Research PR-655) and custom software. We measured CIE XYZ colorimetric coordinates for a range of 8-bit RGB values (0–255) on each monitor and fitted the color characterization model described by Brainard, Pelli, and Robson (2002) in their equations (9) to (17). This model assumes that a gamma function maps RGB framebuffer values to RGB primary intensities and that an affine transformation maps RGB primaries to XYZ color coordinates. We used these mappings to create the stimuli described below, in which luminance varied across pixels and chromaticity was held constant at $x = 0.313$, $y = 0.329$ (the D65 standard). The stimulus luminances that we specify are from the participant's viewing position, that is, they take into account the attenuation of stimuli on the monitors due to the transmittance and reflectance of the beam splitter.

The apparatus also included a printed palette of labeled, achromatic patches that the participant used when reporting the perceived gray of the foreground test patch (Figure 3). The palette was located in front of the chinrest, just below the AR view (Figure 2C), so that the participant could glance back and forth between the palette and the AR display. It showed 14 square patches, each with side length 2.5 cm. The square patches were labeled with their achromatic Munsell values, which ranged from 3.0 to 9.3. To provide context, the

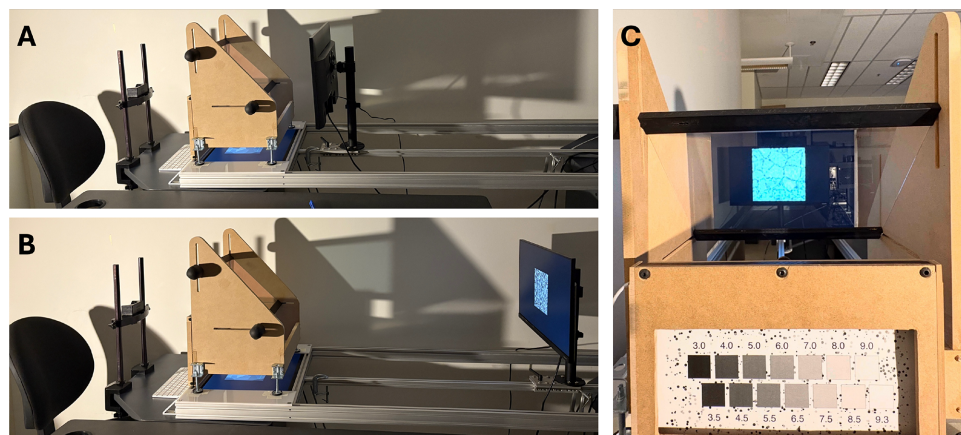


Figure 2. The AR apparatus included an upward-facing foreground monitor (here just visible below the wooden frame), upright background monitor at an adjustable distance, beam splitter (mounted in the wooden frame), chinrest, keyboard, and chair. The lightness palette (Figure 3) was mounted at the bottom of the wooden frame, facing the participant. The apparatus is shown illuminated, as it was during the experiment, by a lamp behind the participant and to their right. (A) In most stimulus conditions, the foreground monitor was at a viewing distance of 73 cm, and the background monitor was at 147 cm. (B) In some conditions, the foreground and background monitors were both at a viewing distance of 73 cm. (C) The participant's view of the display, showing a stimulus from the Eroded condition. This monocular photograph does not convey the clear depth separation between foreground and background that was visible in the actual stimuli. Furthermore, the bluish tint of the stimuli is an artifact of the camera's spectral sensitivity, and in person, the achromatic appearance of the stimuli was well matched to the printed Munsell patches.

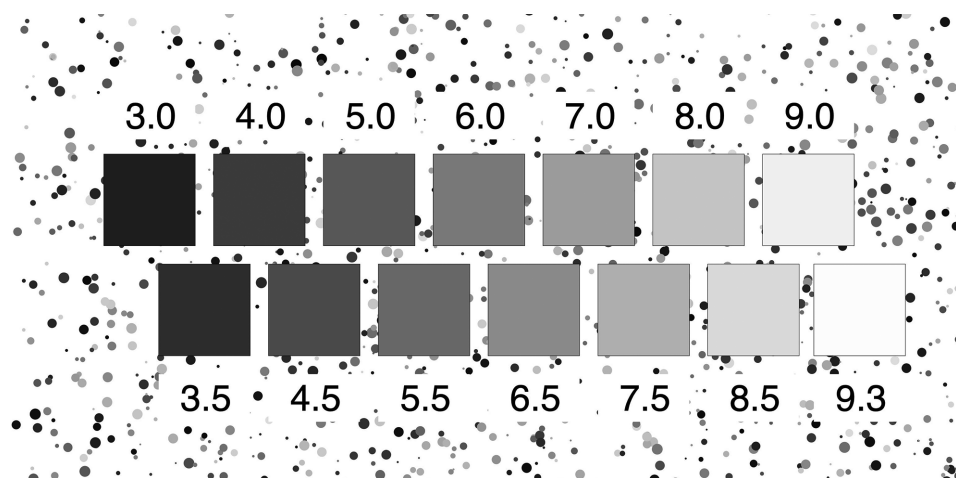


Figure 3. The printed palette of achromatic patches, which participants used to make lightness judgments. Munsell values ranged from 3.0 to 9.3.

squares were surrounded by small achromatic dots with reflectance ranging from 0.03 to 0.90, which provided luminance variations that helped to anchor the range of perceived lightness (Gilchrist, 2006). The palette was printed using an HP Color LaserJet Pro printer (model M452dw). We characterized the printer's mapping from PostScript graylevels to printed reflectances using a photometer (Konika Minolta LS-110) and a 99% reflectance standard (Spectralon SRS-99-020). This allowed us to print the palette with precisely specified reflectances at each point. See [Appendix A](#) for details of the characterization method.

Stimuli

There were eight stimulus conditions (Figure 1). In the Baseline condition, the foreground stimulus was a Voronoi pattern occupying a square region with approximately 25 patches, shown on the upward-facing screen (viewing distance 73 cm). The background stimulus was a finer Voronoi pattern with approximately 1,000 patches, shown on the upright screen (147 cm). Both stimuli subtended 11.0° of visual angle horizontally and vertically. Individual patches in the foreground and background stimuli were assigned random luminances values uniformly distributed between 5 and 50 cd/m^2 . (As explained above, all luminances reported here are measured from the participant's viewing position.) Different, randomly generated Voronoi patterns were shown on each trial. The central patch in the foreground stimulus was designated as the test patch, and participants rated its lightness (details under Procedure). The test patch was randomly assigned one of seven luminances, evenly spaced from 5 to 50 cd/m^2 . An outlined green square, one pixel thick and with a side length 8.6° ,

was rendered in the foreground stimulus, and a similar red square was rendered in the background stimulus. These squares were used to align the foreground and background stimuli from the participant's viewing position (details under Procedure). We did not test whether these outlined squares affected participants' lightness matches, but they were very thin, several degrees of visual angle away from the test patch, and not a salient part of the stimulus, so we think it unlikely that they had a meaningful effect.

Two further conditions used the same foreground Voronoi pattern as the Baseline condition but different background stimuli. (For brevity, here and below, we use the phrase "the same Voronoi pattern" to mean new random samples from the same statistical distribution of Voronoi patterns.) In the Solid-Background condition, the background was a uniform gray square with luminance equal to the mean luminance (27.5 cd/m^2) of the background Voronoi patterns in the Baseline condition. In the No-Background condition, the background monitor uniformly displayed its minimum luminance (0.53 cd/m^2).

Two conditions manipulated the foreground stimulus and used the same Voronoi background as the Baseline condition. In the Eroded condition, the foreground was the same Voronoi pattern as in the Baseline condition, except that pixels within 0.22° of a boundary between patches were assigned the minimum displayable luminance (0.53 cd/m^2). (Hence, the black boundary between patches was 0.44° wide.) In the Solo condition, only the test patch of the foreground Voronoi pattern was shown, and all other patches were assigned the minimum displayable luminance (0.53 cd/m^2).

Finally, three conditions manipulated nonpictorial cues. In the Moving-Background condition, the background stimulus was the same Voronoi pattern as

in the Baseline condition, but it moved in a circular 2D trajectory about the midpoint of the display, with radius 1.1° and period 1.0 s. That is, the image was always upright, and on each frame, its center was shifted to $(x, y) = (R\cos(2\pi t/\tau), R\sin(2\pi t/\tau))$, where $R = 1.1^\circ$, $\tau = 1.0$ s, and t is the time in seconds since stimulus onset. The stimulus was shifted with circular wraparound, so that it always occupied the same square on the monitor, and stimulus regions that were shifted beyond one boundary of the square were wrapped around to the opposite side. (Specifically, we shifted the position using the “cirshift” function in MATLAB.) In the Monocular condition, the stimulus was the same as in the Baseline condition, but participants wore an eye patch over their left eye, eliminating binocular disparity as a cue for distinguishing between foreground and background stimuli. In the Near condition, the stimulus was the same as in the Baseline condition, except that the background screen was at approximately the same viewing distance as the foreground screen (73 cm), and the size of the background stimulus in pixels was reduced to maintain the same size in degrees of visual angle. In this condition, the difference between the foreground and background viewing distances was much less than in the Baseline condition, but they were only approximately matched (e.g., to within 1 cm), so it is likely that a small but nonzero binocular disparity remained between the foreground and background stimuli.

Procedure

The participant sat at a table supporting the AR apparatus, and their head position was stabilized with a chinrest. The experimenter read instructions from a prepared script at the start of the experiment, and before each block of trials.

Before the experiment began, the participant aligned the foreground and background stimuli. With their head in the chinrest, they used arrow keys to adjust the horizontal and vertical position of the foreground stimulus to superimpose its outlined green square onto the red square in the background stimulus. Once the alignment was made, the main experiment began. The green and red squares were shown in the stimuli throughout the experiment and the participant could make further alignment adjustments using the arrow keys if necessary.

On each trial, the stimulus from a randomly chosen condition was shown. The foreground test patch was randomly assigned one of seven test luminances, evenly spaced from 5 to 50 cd/m². At the start of the trial, the outline of the test patch flashed red for 250 ms to indicate to the participant which was the test patch. At any time, the participant could press a key labeled “flash” to make the test patch flash red again as a reminder. Several keys on the keyboard were labeled

with the Munsell ratings 3.0 to 9.3 that were shown on the lightness palette (Figure 3), and the participant pressed one of these keys to report the perceived shade of gray of the test patch. Response time was unlimited. The participant also gave a qualitative rating on each trial. After their lightness response, they pressed one of three labeled keys to indicate whether the lightness judgment was straightforward (labeled “ok”), difficult because the test patch appeared to have more than one lightness value due to interference from the background stimulus (“mix”), or difficult because the test patch had low visibility (“faint”). When reporting the results below, we refer to these ok/mix/faint keypresses as observers’ “qualitative responses”. A short beep acknowledged each keypress response. After the two responses, both monitors were blank for 500 ms, and then the next trial began.

There were eight conditions, and each of the seven test patch luminances was shown on eight trials, for a total of 448 trials. Trials were divided into four blocks. Blocks 1 and 2 (168 trials each) showed the Baseline, Solid-Background, No-Background, Eroded, and Solo conditions. Block 3 (56 trials) showed the Monocular condition, and Block 4 (56 trials) showed the Near condition. Trial order was randomized within blocks, and block order was counterbalanced across participants. A short break was provided between blocks to mitigate fatigue. New, randomly generated Voronoi patterns were shown on each trial. Before the start of each block, the experimenter gave a brief reminder of the instructions. The full experimental session took approximately 45 minutes.

Results

Experiment 1 included eight conditions that addressed three related questions. For clarity, we group the results according to these three goals: effects of background context, effects of foreground context, and effects of segmentation and depth cues. Trial-by-trial data and code for analyses are available at <https://osf.io/sydwf>.

Effects of background context

We first examined how altering the background plane affected reflectance matches. Figure 4A shows results for the Baseline, Solid-Background, and No-Background conditions, plotting the mean reflectance match across participants as a function of test patch luminance. In this plot and later plots, we have converted the achromatic Munsell values that participants used to respond into reflectance values (see Appendix B for the conversion formula). In all conditions, match reflectance was an approximately linear function of test patch luminance, a notable regularity that we return to

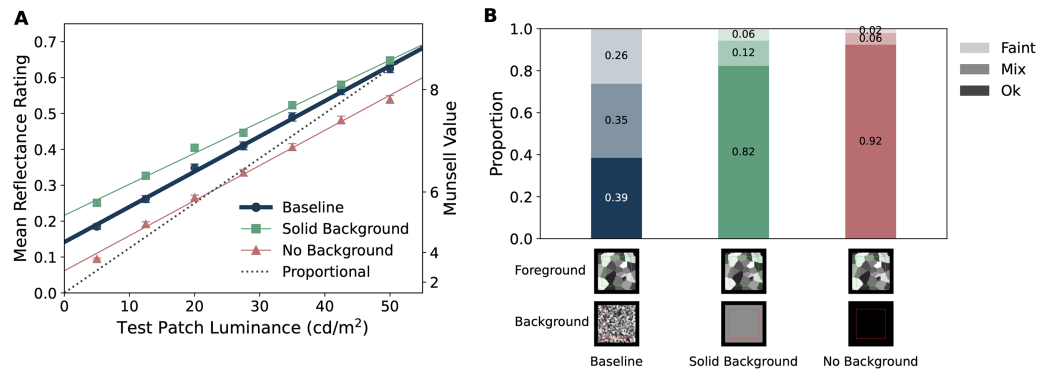


Figure 4. Results from Baseline, Solid-Background, and No-Background conditions. **(A)** Mean reflectance ratings as a function of test patch luminance. **(B)** Qualitative judgments on the ease of making reflectance ratings.

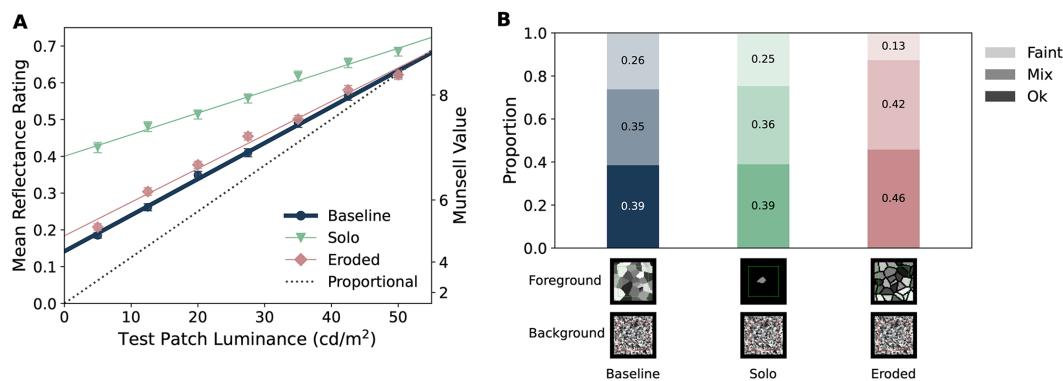


Figure 5. Results from Baseline, Eroded, and Solo conditions. **(A)** Mean reflectance ratings as a function of test patch luminance. **(B)** Qualitative judgments on the ease of making reflectance ratings.

in the General Discussion. The most prominent effect of having different backgrounds in these three conditions was a vertical shift in reflectance ratings: Compared to the Baseline, matches in the Solid-Background condition were slightly higher, and matches in the No-Background condition were lower.

To quantify these effects, we fit the following linear model to participants' reflectance matches.

$$r = m(l - l_0) + b \quad (1)$$

Here, l is test patch luminance (the independent variable), and r is the mean reflectance rating (the dependent variable). The parameter m is the fitted slope, which captures how widely reflectance ratings vary as a function of luminance. The parameter b is the fitted y -intercept at the vertical line $l = l_0 = 27.5$ cd/m². Here, l_0 is the mean of the seven evenly spaced test patch luminances (a constant), and as a result, the fitted intercept parameter b is equal to the mean reflectance rating across all test patch luminances (see Appendix C). With this parameterization, the fitted

parameters describe simple and meaningful features of the data: m indicates the range of reflectance ratings, and b indicates the mean reflectance rating over all test patch luminances.

The colored lines in Figure 4A show least squares fits of this linear model to data pooled across participants in each condition. For reference, Figure 4A also shows a dotted line ("proportional" in the legend) that passes through the origin and through the mean reflectance rating at the highest test patch luminance, averaged over all participants and conditions in Experiment 1. This line represents a simple proportional mapping between luminance and perceived lightness. The same dotted line is shown in all plots for Experiment 1 (Figures 4, 5, 6).

To compare results across conditions, we fit Equation 1 separately to data from each participant in each condition and carried out paired-sample t -tests on the fitted parameters m and b . Slope was significantly lower in the Solid-Background condition than in the Baseline condition ($\Delta m = -0.0012 \pm 0.0002$, $t(11) =$

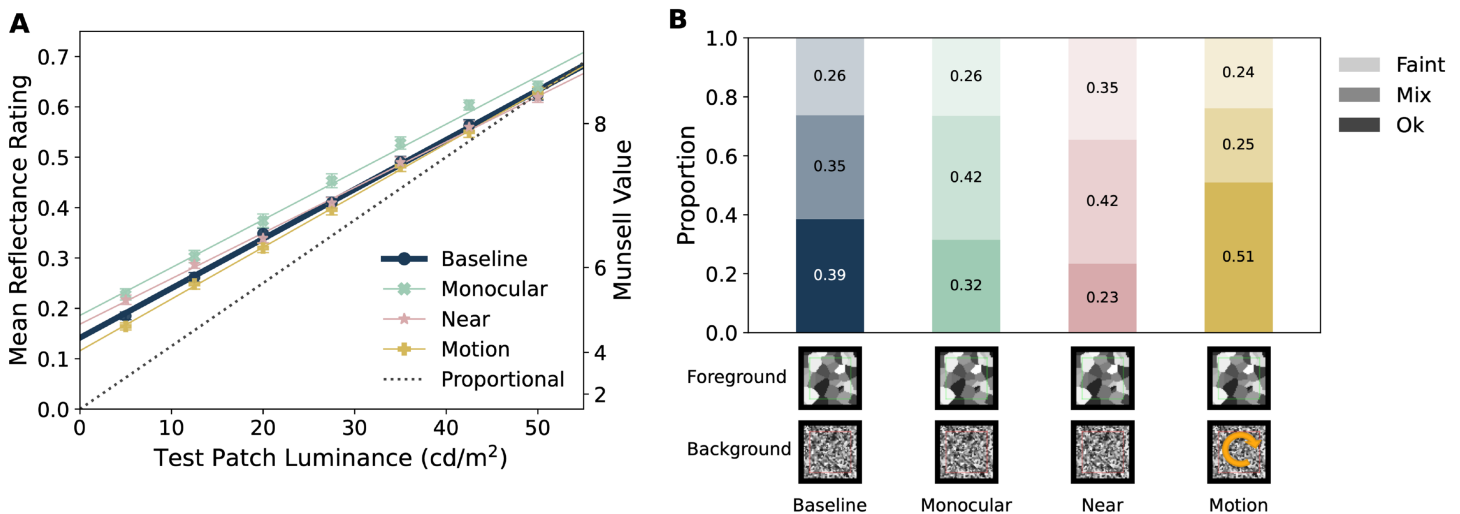


Figure 6. Results from Baseline, Motion, Monocular, and Near conditions. **(A)** Mean reflectance ratings as a function of test patch luminance. **(B)** Qualitative judgments on the ease of making reflectance ratings.

−4.86, $p < 0.001$), and the intercept was significantly higher ($\Delta b = 0.042 \pm 0.003$, $t(11) = 11.59$, $p < 0.001$), resulting from an overall upward shift in perceived reflectance. Slope was not significantly different in the No-Background condition than in Baseline ($\Delta m = -0.00005 \pm 0.00026$, $t(11) = -0.177$, $p = 0.86$), and the intercept was significantly lower in the No-Background condition ($\Delta b = -0.081 \pm 0.004$, $t(11) = -20.57$, $p < 0.001$), indicating lower perceived reflectance in the No-Background condition. The intervals ($\pm$) that we report in these tests are standard errors. The Supplementary Information provides data and code for all analyses reported here.

We report line fits in reflectance units, but it may be helpful to illustrate how these translate into Munsell values. In Figure 4A, the intercept parameter for the No-Background condition is $b = 0.331$, which corresponds to a Munsell value of 6.4. The intercept for the Solid-Background condition is $b = 0.454$, corresponding to a Munsell value of 7.3. Thus, observers' mean lightness matches were approximately 0.9 Munsell units higher in the Solid-Background condition than in the No-Background condition, a difference that is clearly visible as a vertical shift between the two fitted lines in Figure 4A.

There were also notable differences in qualitative responses across these three conditions (Figure 4B). In the Baseline condition, just 39% of participants' reflectance ratings were judged "ok," and the rest were difficult to make because the test patch appeared to have more than one reflectance or was too faint. In the Solid-Background and No-Background conditions, most ratings were judged "ok". Chi-square tests indicated that qualitative responses were independent of luminance in these three conditions (all $p > 0.14$).

Effects of foreground context

We next examined how manipulating foreground context altered lightness perception. Figure 5A shows results for the Baseline, Eroded, and Solo conditions. The Baseline results shown here are identical to those in Figure 4A and are included here for comparison. Slope was significantly different in the Eroded condition than in the Baseline condition ($\Delta m = -0.0007 \pm 0.0003$, $t(11) = -2.45$, $p = 0.032$) and the intercept was slightly higher ($\Delta b = 0.023 \pm 0.004$, $t(11) = 5.49$, $p < 0.0002$), but this effect was small. Slope was markedly lower in the Solo condition than in the Baseline condition ($\Delta m = -0.0033 \pm 0.0003$, $t(11) = -10.61$, $p < 0.001$), and the intercept was markedly higher ($\Delta b = 0.127 \pm 0.004$, $t(11) = 27.45$, $p < 0.001$); these are the largest effects of any that we tested for in this experiment.

Qualitative responses (Figure 5B) were broadly similar across these conditions. In all cases, less than half the reflectance ratings were judged "ok", and others were difficult because the test patch appeared mixed or faint. Chi-square tests revealed that qualitative responses were independent of luminance in the Eroded condition ($p = 0.33$) but showed a significant association in the Solo condition ($\chi^2(12) = 124.79$, $p < 0.001$), with "faint" responses more common at low luminances, suggesting the test patch was particularly difficult to judge when appearing alone against a dark surround.

Effects of nonpictorial cues

Finally, we tested whether motion and disparity cues influenced perceived reflectance. Figure 6A shows results for the Baseline, Motion, Monocular, and Near conditions. (Again, the Baseline data are the same as

that reported in the previous two figures.) Overall, there were no large differences between these conditions, although some statistical comparisons did achieve significance. Slope was not significantly different in the Motion condition than in the Baseline condition ($\Delta m = 0.0004 \pm 0.0003$, $t(11) = 1.75$, $p = 0.10$), but the intercept was significantly lower ($\Delta b = -0.014 \pm 0.003$, $t(11) = -3.55$, $p < 0.01$). Slope was not significantly different in the Monocular condition than in the Baseline condition ($\Delta m = -0.0003 \pm 0.0003$, $t(11) = -1.12$, $p = 0.29$), but the intercept was significantly higher ($\Delta b = 0.035 \pm 0.004$, $t(11) = 7.85$, $p < 0.001$). Slope was lower in the Near condition than in the Baseline condition ($\Delta m = -0.0008 \pm 0.0003$, $t(11) = -2.59$, $p = 0.03$), but the intercept was not significantly different ($\Delta b = 0.006 \pm 0.005$, $t(11) = 1.23$, $p = 0.24$).

Qualitative responses were similar across conditions, the most notable difference being that just over 50% of reflectance ratings were judged “ok” in the Motion condition, compared to less than 40% in the other three conditions (Figure 6B). Chi-square tests confirmed that qualitative responses were independent of luminance level across all three conditions (all $p > 0.56$).

Discussion

In this experiment, lightness matches were most strongly influenced by changes to the foreground layer, particularly by those that reduced the context around the test patch. See Figure 7 for a summary of the results. In the Baseline condition, reflectance ratings varied linearly with test patch luminance and spanned a broad range. When the surrounding foreground elements were removed in the Solo condition, reflectance ratings were higher and varied less with test patch luminance. Surprisingly, even large background manipulations—such as replacing the background with uniform black or gray fields (No-Background and Solid-Background conditions), adding motion (Moving-Background), or collapsing the depth separation between layers (Near)—produced only modest shifts in reflectance ratings. These findings highlight the key role of foreground context in supporting lightness perception in AR.

These results also vividly illustrate some important features of lightness perception in AR. We can see three such features in the results of the Baseline, No-Background, and Solid-Background conditions (Figure 4A). First, there is bleed-through from the background when observers judge reflectance in the foreground. In the No-Background condition, where background luminance was minimal, the intercept parameter (i.e., the mean reflectance rating) was $b = 0.331$. In the Solid-Background condition, where background luminance was 27.5 cd/m^2 , the intercept parameter was 0.123 reflectance units higher ($b =$

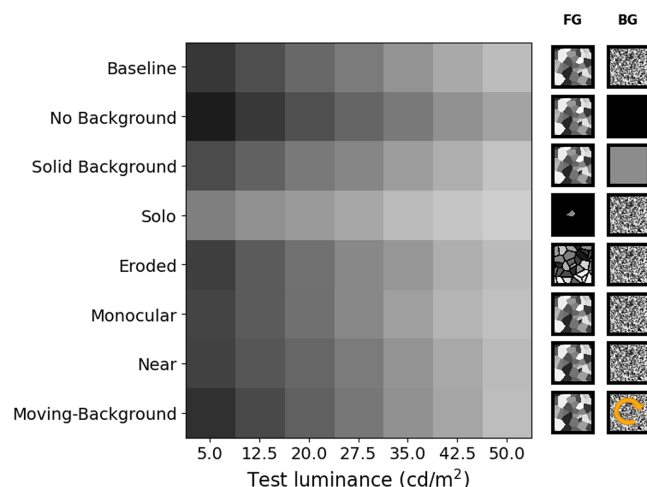


Figure 7. Illustration of mean perceived reflectance across conditions in Experiment 1. Each row corresponds to one condition, and each column corresponds to one of the seven test luminances. Each square represents the mean perceived reflectance in that condition and luminance, allowing a visual comparison of how perceived lightness varied with luminance across conditions. For example, the Solo row shows that perceived reflectances were markedly higher in this condition. Thumbnails on the right show examples of the foreground (FG) and background (BG) in each condition.

0.454). Thus, the solid background luminance had a marked effect on reflectance ratings in the foreground.

Second, there is only *partial* bleed-through from the background. Continuing the analysis from the previous paragraph, how much of an increment in the luminance of the foreground test patch would be needed in the No-Background condition to produce the same increase in reflectance ratings observed in the Solid-Background condition? The slope of the fitted line in the No-Background condition was $m = 0.00978$, so an increase of 0.123 reflectance units would require an increase of $0.123/0.00978 = 12.6 \text{ cd/m}^2$ in the foreground test patch luminance. This is less than half of the 27.5 cd/m^2 increase in background luminance that produced the same increase of 0.123 in reflectance ratings. Thus, foreground luminance is much more effective than background luminance at increasing reflectance ratings. Interestingly, this is a purely perceptual effect. If observers simply mapped local retinal luminance at the test patch location to an estimate of reflectance, then it would not matter whether a luminance increment came from the foreground or background stimulus. Here, though, foreground luminance had more than twice the impact of background luminance on perceived reflectance.

Third, the amount of partial bleed-through depends on the strength of scission cues between foreground and background. The mean reflectance rating (i.e., the fitted intercept b) was 0.080 higher in the Baseline condition

($b = 0.411$) than in the No-Background condition ($b = 0.331$) and 0.123 higher in the Solid-Background condition ($b = 0.454$) than in the No-Background condition ($b = 0.331$). On this measure, bleed-through from the background was $0.123 / 0.080 = 1.54$ times greater in the Solid-Background condition than in the Baseline condition, even though these two conditions had the same mean background luminance. The Baseline condition had a textured Voronoi background that provided more edges and disparity cues, evidently resulting in greater scission, and less disruption from the background. This too is a purely perceptual effect.

We also find it noteworthy that the slope of the fitted line was much lower in the Solo condition ($m = 0.00588$) than in all other conditions (average $\bar{m} = 0.00946$; Figure 5A). Not only was perceived reflectance higher in the Solo condition, but it also varied much less with luminance. This recalls the Gelb effect, where an isolated test patch in an otherwise dark scene appears light gray or white over a wide range of luminances (Gelb, 1938). This suggests that in addition to the well-known phenomenon of bleed-through from the background, another reason for the reduced range of lightness perception in AR may often be a lack of accompanying, contextual elements in the foreground plane.

Experiment 2

In Experiment 2, we introduced new conditions that addressed three questions raised by the results of Experiment 1.

First, in Experiment 1, the Solo condition had a much larger effect on lightness matches than the Eroded condition, when both were compared to the Baseline condition. This difference could reflect either the presence of multiple foreground patches in the Eroded condition or the variation in luminance between those patches. To dissociate these factors, we created the Eroded-Uniform condition (Figure 8, top row), which used the same modified Voronoi foreground pattern as the Eroded condition, but assigned the same luminance to all patches.

Second, a surprising result in Experiment 1 was that lightness judgments were not markedly different in the Baseline and Motion conditions, even though motion can often be a strong grouping cue. We created the Solo-Motion condition (Figure 8, middle row) to test whether the weak effect of motion was due to a ceiling effect: Perhaps motion did not have a large effect on lightness judgments because scission between the foreground and background was already strong in the Baseline condition. The Solo-Motion condition combined the foreground from the Solo condition, which had produced only a narrow range of lightness judgments, with the background from the Motion

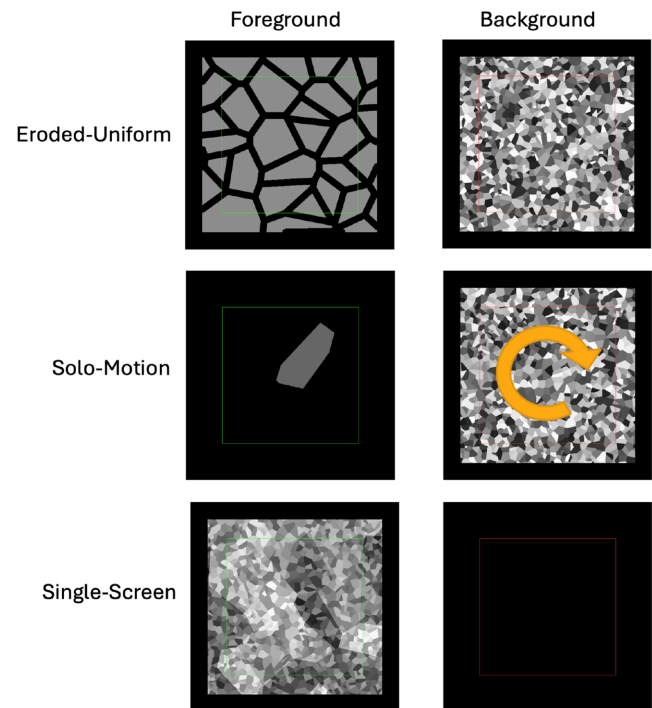


Figure 8. Examples of stimuli in the new conditions of Experiment 2. Eroded-Uniform, same as Uniform, except that all foreground patches had the same luminance. Solo-Motion, same as Solo, except that the background moved in a circular trajectory as in the Motion condition. Single-Screen, same as Baseline, except that the foreground and background stimuli were summed and shown on the foreground monitor, and the background monitor was set to black.

condition. This provided a test of whether motion could widen the range of perceived reflectance, starting from a condition (Solo) with a narrower lightness range than the Baseline condition.

Third, the Single-Screen condition (Figure 8, bottom row) tested whether the relatively small effect of reducing the depth difference between the foreground and background planes in the Near condition of Experiment 1 was due to residual disparity between the foreground and background layers. In this condition, we displayed both layers on the foreground screen to eliminate the depth difference altogether. This allowed us to assess whether background interference increases when the foreground and background stimuli are coplanar.

Methods

Participants

Thirteen new, naive participants (8 female, 5 male; mean age 22.5 years, range 18–30 years) were recruited from York University. In this experiment, we screened

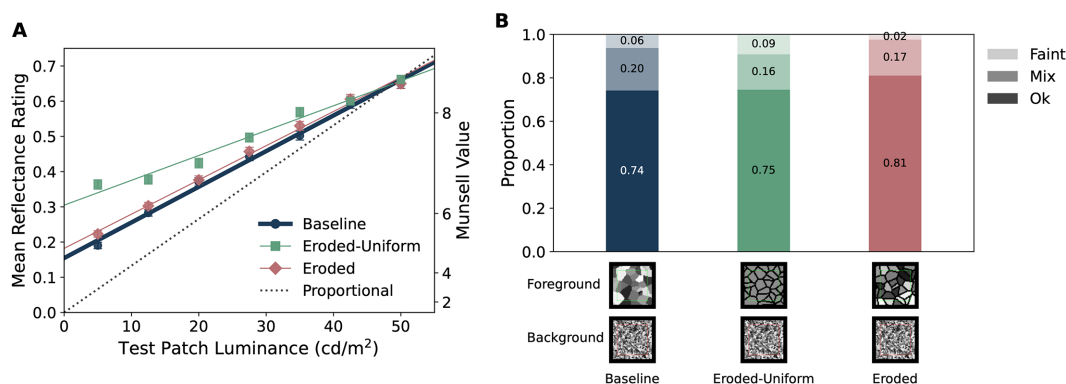


Figure 9. Results from Baseline, Eroded, and Eroded-Uniform conditions. **(A)** Mean reflectance ratings as a function of test patch luminance. **(B)** Qualitative judgments on the ease of making reflectance ratings.

participants using the Randot Stereo Test (Stereo Optical Company, Chicago, IL, USA) and required a stereo acuity threshold of 40 arcseconds or better. All other recruitment, consent, and compensation procedures were the same as in [Experiment 1](#).

Data from one participant was excluded from analysis because they completed the experiment very quickly (3 minutes, compared to the average of 10 minutes for other participants), and their responses were highly variable. We show this participant's results in the Supplementary Information.

Stimuli

The experiment included the Baseline, Eroded, Solo, and Near conditions from [Experiment 1](#) and the following three new conditions ([Figure 8](#)). The Eroded-Uniform condition was the same as the Eroded condition, except that all patches in the foreground Voronoi pattern were assigned the same luminance as the test patch. The Solo-Motion condition showed the isolated test patch from the foreground of the Solo condition and the moving background of the Moving-Background condition. The Single-Screen condition showed the foreground and background stimuli of the Baseline condition, with their luminances summed and displayed on the foreground monitor, and the background monitor was set to its minimum luminance ($0.53 \text{ cd}/\text{m}^2$).

Procedure

The procedure was the same as in [Experiment 1](#), except that each test patch luminance was shown nine times instead of eight. The Baseline, Eroded, Eroded-Uniform, Solo, and Solo-Motion conditions were shown in Blocks 1 to 3 (105 trials each), and the Near and Single-Screen conditions were shown in Block

4 (126 trials). Block order was counterbalanced across participants.

Results

Foreground luminance range

We first examined whether luminance variation within the foreground texture was necessary for maintaining a broad range of perceived reflectance. [Figure 9A](#) shows results for the Baseline, Eroded, and Eroded-Uniform conditions. In all plots of results in Experiment 2, the dotted “proportional” line passes through the origin and the mean of all reflectance ratings at the highest test patch luminance across all conditions of Experiment 2. As in Experiment 1, mean reflectance matches were an approximately linear function of test patch luminance. Slope was not significantly different in the Eroded condition than in the Baseline condition ($\Delta m = -0.0004 \pm 0.0003$, $t(11) = -1.35$, $p = 0.21$), and the intercept was slightly lower ($\Delta b = 0.016 \pm 0.004$, $t(11) = 3.65$, $p = 0.004$). However, slope was significantly lower in the Eroded-Uniform condition than in the Baseline condition ($\Delta m = -0.0031 \pm 0.0003$, $t(11) = -9.20$, $p < 0.001$), and the intercept was appreciably higher ($\Delta b = 0.066 \pm 0.005$, $t(11) = 13.19$, $p < 0.001$). These findings show that luminance variation in the Eroded condition played an important role in maintaining a wide range of perceived reflectance.

Qualitative responses were broadly similar across these conditions ([Figure 9B](#)). Chi-square tests showed that qualitative responses were independent of luminance in the Eroded condition ($p = 0.48$) but showed a significant dependence in the Eroded-Uniform condition ($\chi^2(12) = 152.77$, $p < 0.001$), with “faint” responses more frequent at low luminances.

The proportion of “ok” responses was much higher in this experiment (around 75%–80%) than

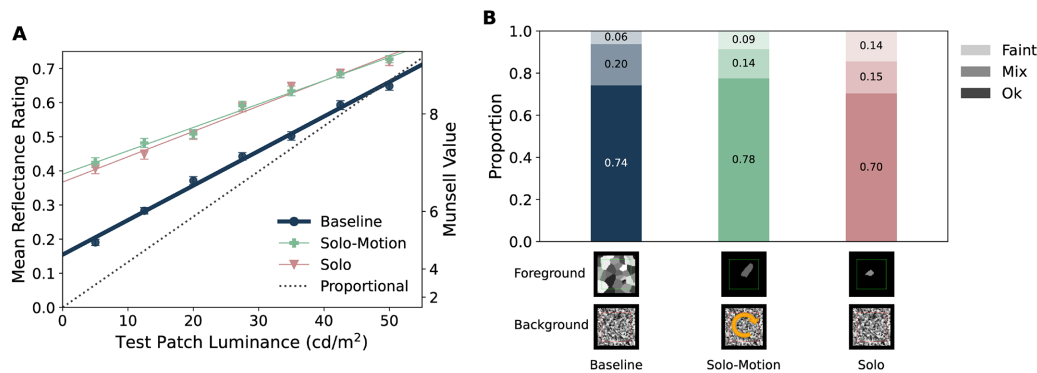


Figure 10. Results from Baseline, Solo, and Solo-Motion conditions. (A) Mean reflectance ratings as a function of test patch luminance. (B) Qualitative judgments on the ease of making reflectance ratings.

in Experiment 1 (around 40%), even in the Baseline and Eroded conditions that were identical in the two experiments. Individual differences may have played a role. We also speculate that the proportion of “ok” responses was affected by the range of interleaved conditions in each experiment. Experiment 2 included the Single-Screen condition, where there were few cues supporting scission between the foreground and background stimuli, and this may have led participants to report other conditions as being relatively easy. Consistent with this explanation, we found that on the subset of trials in the Baseline condition that were run before the Single-Screen condition, there were 66% “ok” responses, whereas on Baseline trials after the Single-Screen condition, there were 82% “ok” responses. Thus, at least part of the difference between qualitative responses in the two experiments seems to have been due to the differences in the interleaved conditions.

Motion and foreground context

Next we tested whether background motion could have a large effect on the range of perceived lightness in a condition with little foreground context. Figure 10A shows results for the Baseline, Solo, and Solo-Motion conditions. Slope was significantly lower in the Solo condition than in the Baseline condition ($\Delta m = -0.0027 \pm 0.0003$, $t(11) = -7.82$, $p < 0.001$), and the intercept was significantly higher ($\Delta b = 0.138 \pm 0.005$, $t(11) = 26.71$, $p < 0.001$), consistent with the same comparison in Experiment 1. Similarly, slope was lower in the Solo-Motion condition than in the Baseline condition ($\Delta m = -0.0033 \pm 0.0003$, $t(11) = -9.43$, $p < 0.001$), and the intercept was significantly higher ($\Delta b = 0.146 \pm 0.005$, $t(11) = 28.02$, $p < 0.001$). Results were practically identical in the Solo and Solo-Motion conditions, indicating that background motion did not appreciably widen the range of perceived reflectance.

Qualitative responses were similar across conditions, with high rates of “ok” responses in all three cases (Figure 10B). Chi-square tests revealed significant associations between luminance and qualitative response in both the Solo ($\chi^2(12) = 189.84$, $p < 0.001$) and Solo-Motion ($\chi^2(12) = 144.67$, $p < 0.001$) conditions, with “faint” responses more common at low luminances. This pattern was absent in the Baseline condition ($p = 0.30$).

Depth difference

Finally, we examined whether eliminating depth separation entirely would increase perceptual bleed-through from the background. Figure 11A shows results for the Baseline, Near, and Single-Screen conditions. Results were broadly similar across these conditions. Slope was not significantly different in the Near and Baseline conditions ($\Delta m = -0.0005 \pm 0.0003$, $t(11) = -1.81$, $p = 0.097$), nor was the intercept ($\Delta b = -0.036 \pm 0.018$, $t(11) = -1.97$, $p = 0.074$). In the Single-Screen condition, slope was significantly lower than in the Baseline condition ($\Delta m = -0.0009 \pm 0.0004$, $t(11) = -2.32$, $p = 0.041$), and the intercept was also significantly lower ($\Delta b = -0.015 \pm 0.006$, $t(11) = -2.40$, $p = 0.036$). While these differences reached statistical significance, they were small, and overall, the manipulation of viewing distance (Near and Single-Screen) did not produce large changes in perceived lightness.

However, there were differences in qualitative responses. In the Baseline and Near conditions, the proportion of “ok” responses was high, as in previous conditions. In the Single-Screen condition, it was much lower, even though the reflectance ratings themselves were not substantially different. Chi-square tests confirmed that qualitative responses were independent of luminance in both the Near and Single-Screen conditions ($p = 0.36$ and $p = 0.92$, respectively).

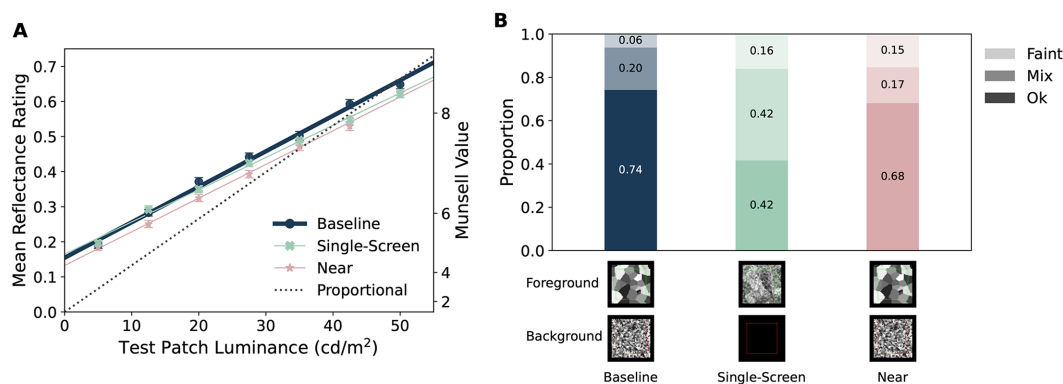


Figure 11. Results from Baseline, Near, and Single-Screen conditions. **(A)** Mean reflectance ratings as a function of test patch luminance. **(B)** Qualitative judgments on the ease of making reflectance ratings.

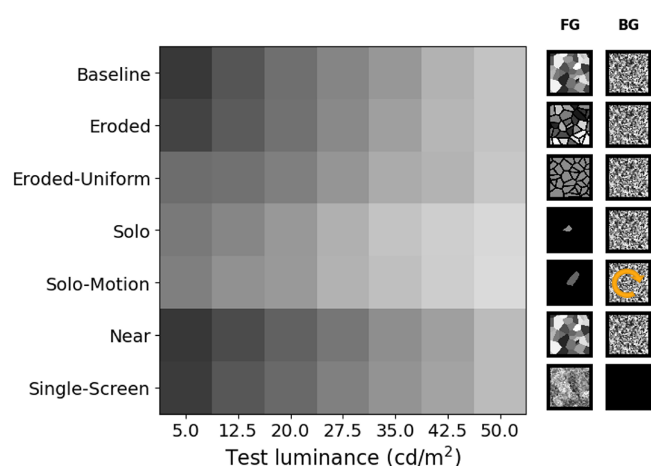


Figure 12. Illustration of mean perceived reflectance for each condition (rows) and test luminance (columns) in Experiment 2.

Discussion

Experiment 2 provided clear answers to the questions that motivated its three new conditions (see Figure 12 for summary of results). First, in the Eroded-Uniform condition, perceived reflectance was compressed compared to the Eroded condition: Slope decreased, and intercept increased, much as in the Solo condition of Experiment 1. This indicates that luminance variation in the foreground was critical for maintaining a wide range of perceived lightness.

Second, lightness matches in the Solo-Motion condition were nearly identical to those in the Solo condition, indicating that the lack of an effect of motion in Experiment 1 was not just a ceiling effect. This result also suggests that the reduced lightness range in the Solo condition (i.e., low slope m), which persisted in the Solo-Motion condition, was not the result of inadequate cues to scission but rather resulted from some other factor, such as the lack of foreground context.

Finally, in the Single-Screen condition, removing all depth separation did not have any effect on slope and intercept compared to the Baseline condition. However, the lower proportion of “ok” responses in the qualitative judgments suggests that separation in depth contributed to the quality of lightness percepts.

Together, these findings show that while motion and disparity can be important cues for scission, they do not always play a decisive role in lightness perception. Instead, strong foreground structure—particularly luminance variation within textured regions—plays a central role in maintaining a wide range of perceived lightness in OST-AR.

General discussion

These experiments investigated how foreground and background context, motion, and depth affect lightness perception in OST-AR. We found that changes to context—especially in the foreground—had clear effects on how light or dark a virtual patch appeared. Participants’ lightness judgments followed test luminance in an approximately linear fashion, and the slope and intercept of this linear mapping varied across conditions. These shifts revealed that several stimulus properties affected either the overall lightness of the patch (i.e., making it consistently lighter or darker) or how strongly its appearance changed with luminance. In particular, removing foreground context (Solo condition) had the largest impact, compressing the range of perceived reflectance and elevating lightness judgments. This suggests that strong visual grouping and local context are key to generating a wide range of lightness percepts in augmented reality. This is an encouraging finding for practical applications of AR, since we usually have much more control over the computer-generated foreground of the display than over the real-world background.

How visual context shapes lightness perception

Across the two experiments, the clearest finding was that changes to the foreground had the strongest impact on perceived lightness. When the test patch was shown in a rich foreground texture—especially one with a wide range of luminance values, as in the Baseline and Eroded conditions—perceived reflectance spanned a wide range. When the patch was shown alone or with minimal context (Solo and Eroded-Uniform conditions), lightness ratings spanned a smaller range and were shifted upward. Changes to the background, such as making the background uniform (Solid-Background) or removing it entirely (No-Background), produced smaller effects. Background manipulations tended to shift lightness judgments upward or downward, as if the observer counted some portion of the background luminance as belonging to the foreground, and the amount depended on the scission cues available in each condition.

These findings are consistent with some current theoretical frameworks for understanding lightness perception. [Adelson \(2000\)](#) noted that the physical mapping from surface reflectance to stimulus luminance (the “atmospheric transfer function”) is usually linear, even in the presence of complex phenomena such as haze and interreflections. He suggested that as a result, the inverse perceptual mapping from stimulus luminance to perceived reflectance (the “lightness transfer function”) might also sometimes be linear, although he emphasized that this is an empirical question. Here we found that the lightness transfer function was in fact linear, which is broadly consistent with previous studies (e.g., [Zeiner & Maertens, 2014](#)), although departures from nonlinearity have also been noted ([Allred, Radonjic, Gilchrist, & Brainard, 2012](#)). This was true even though participants made responses on the Munsell value scale, which is nonlinearly related to reflectance and luminance, and so the observed linearity is unlikely to be a by-product of the response scale.

Our findings are also broadly consistent with [Gilchrist’s \(2006\)](#) anchoring theory of lightness perception. The largest effect in the two experiments was that when we removed the foreground context, in the Solo condition, the perceived reflectance of the test patch became much higher and varied less with luminance. The basic claim of anchoring theory is that lightness is calculated relative to the highest luminance (the anchor) in a lighting region: Luminances equal to the anchoring luminance are seen as white, and lower luminances are seen as shades of gray, or black. In the Solo condition, the test patch was the only element in the foreground plane and hence the anchor for that plane. To a first approximation, anchoring theory predicts that such a test patch should tend to appear white, and in fact, the test patch lightness was

high and varied relatively little across luminances. However, anchoring theory also posits that the perceived reflectance of a test patch not only depends on the anchor in its own lighting region but also is also co-determined by other lighting regions in the scene. This aspect of the theory accounts, at least qualitatively, of our finding that the perceived reflectance of the Solo test patch was not completely constant and did vary with luminance.

[Murdoch \(2025\)](#) also found that a high-luminance white frame around a stimulus served as a lightness anchor and improved the perception of dark skin tones. From this study and our own, it is unclear whether a high-luminance anchor alone supports lightness perception or whether having a wide range of luminances, as in most of our foreground stimuli, has additional benefits. This would be straightforward to test with further study. Furthermore, [Murdoch \(2020\)](#) developed a color appearance model of AR stimuli, in which appearance is a weighted sum of foreground and background stimuli, and the weights are determined by scission cues. He measured color appearance in several conditions with various cues to scission and found that the weighted-sum model could accommodate the results. An important difference with the present study is that Murdoch examined the appearance of background stimuli in the presence of veiling foreground stimuli, whereas we examined the appearance of foreground stimuli. Nevertheless, Murdoch’s weighted-sum model is consistent with our finding that perceived reflectance is a linear function of foreground and background luminance, with slope and intercept that vary across conditions with different scission and grouping cues.

Interestingly, the range of perceived reflectance in these experiments did not show a clear relationship to qualitative responses. For example, perceived lightness differed dramatically in the Solo and Baseline conditions of [Experiment 1](#), but participants’ rates of qualitative responses (“ok”, “mix”, or “faint”) were roughly the same. Perceived lightness was also roughly the same in the Baseline and Single-Screen conditions of [Experiment 2](#), but participants had many more “faint” and “mix” responses in the latter condition. Thus, the difficulty of making lightness judgments seems to be dissociable from the range of perceived lightness.

However, when we examined whether qualitative responses varied with luminance level, a consistent pattern emerged: In conditions with rich foreground context, the difficulty of making lightness judgments was unrelated to test patch luminance (all $p > 0.14$). In contrast, conditions with reduced or absent foreground context—the Solo, Solo-Motion, and Eroded-Uniform conditions—showed significant associations between luminance and qualitative response (all $\chi^2(12) > 124$, all $p < 0.001$), with “faint” responses disproportionately

common at low luminances. This suggests that while the overall difficulty of lightness judgments is dissociable from the range of perceived lightness, the visibility of dark surfaces in particular depends critically on the presence of rich foreground context.

Overall, our findings point to the importance of having a rich visual context in the foreground plane for lightness perception. In particular, both the presence of multiple elements and having high-luminance elements are crucial for creating a wide range of perceived lightness in AR.

Motion and depth cues

While foreground manipulations had the strongest effects, motion and depth cues also played a small role. We expected the Motion condition, which introduced relative motion between background and foreground, to enhance scission and to substantially reduce the effect of the background on foreground lightness judgments. However, lightness judgments in the Motion condition were only slightly different from those in the Baseline condition. This may have been because scission cues were already sufficiently strong in the Baseline condition, so adding a motion cue had little additional effect. This speculation is supported by the similarity of the results in the Solo and Solo-Motion condition: There too, even though perceived reflectance spanned a limited range in the Solo condition, introducing motion had little effect, which suggests that the limited range of lightness in those conditions was due to the lack of foreground context and not weak scission between foreground and background.

Binocular depth cues had similarly limited effects. Removing disparity (Monocular condition) or reducing depth separation (Near condition) produced small but statistically significant shifts in lightness judgments. Notably, the Near condition led to fewer “ok” ratings and more “faint” responses, suggesting that disparity cues, without other supports, may sometimes be more important for clarity than for reducing perceptual bleed-through from the background.

Even comparisons that we might have expected to show strong effects—such as Background-Motion versus Background-Solid—showed smaller effects than those involving foreground structure. This reinforces a central finding: The most influential features for lightness constancy are often those within the foreground, not the background.

Implications

The results of this study have practical implications for OST-AR design. A persistent challenge in AR is rendering dark surfaces convincingly. Because of

additive blending, even virtual elements rendered at low luminances allow background light to show through, making them appear washed out or semi-transparent (Peck et al., 2022). Our findings reflect this limitation: Conditions with weak contextual support, such as the Solo condition, produced large upward shifts in lightness. This can undermine realism and also make dark-skinned avatars appear less visible than light-skinned ones, potentially affecting impressions of presence or humanness (Peck et al., 2022).

Our data point to some ways designers can address these challenges. First, strong foreground structure is essential. Foreground textures that include many elements with varied luminance values, as opposed to uniform patterns or isolated patches, are better able to support a broad range of perceived lightness. Second, while background cues such as motion and disparity may help under certain conditions, their effects are relatively small. From a design perspective, this is encouraging: It means that the stimulus properties we can control—those in the foreground—are also the ones that are most effective. However, the present study used a custom-built beam splitter apparatus to achieve precise stimulus control. Whether the effects we report generalize to consumer AR headsets based on waveguide optics, which can present artifacts such as chromatic aberration, and to more complex, naturalistic backgrounds, remain important open questions.

Beyond improving visibility of low-reflectance content, these findings show how perceptual principles can be leveraged to extend the perceptual range of AR displays. In conditions that provided context and scission cues, participants perceived a wide range of lightness. When those cues were degraded or absent, judgments were compressed and biased. This pattern suggests that incorporating perceptual cues—like contrast variation, grouping, and depth separation between planes (including binocular disparity)—can help virtual objects appear more realistic and stable, in a range of lighting and background conditions. Incorporating these and similar insights will be key to making OST-AR systems perceptually robust and socially inclusive. Designers can structure virtual scenes to align with how the visual system interprets lightness, using mechanisms such as local contrast and perceptual organization to support realism and reduce bias.

Keywords: lightness perception, augmented reality, lightness constancy, contextual vision, optical see-through displays

Acknowledgments

This work was initiated while author MK was at Meta Platforms, Inc. Meta did not provide any

proprietary intellectual property or financial support specific to the project beyond providing full-time employment to author MK.

Funded by an NSERC Discovery grant to RFM.

Commercial relationships: none.

Corresponding author: Tasfia Ahsan.

Email: tasfia.ahsan@gmail.com.

Address: Department of Psychology and Centre for Vision Research, York University, Toronto, Ontario, Canada.

References

- Adelson, E. H. (2000). Lightness perception and lightness illusions. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (p. 339–351). Cambridge, MA: The MIT Press.
- Allred, S. R., Radonjic, 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, 1–16, <https://doi.org/10.1167/12.2.7>.
- Anderson, B. L., & Winawer, J. (2005). Image segmentation and lightness perception. *Nature*, 434(7029), 79–83.
- Anderson, B. L., & Winawer, J. (2008). Layered image representations and the computation of surface lightness. *Journal of Vision*, 8(7), 22, <https://doi.org/10.1167/8.7.18>.
- Brainard, D. H., & Maloney, L. T. (2011). Surface color perception and equivalent illumination models. *Journal of Vision*, 11(5), 1–18, <https://doi.org/10.1167/11.5.1>.
- Brainard, D. H., Pelli, D. G., & Robson, T. (2002). Display characterization. In J. Hornak (Ed.), *Encyclopedia of imaging science and technology* (p. 172–188). Hoboken, NJ: Wiley.
- Fairchild, M. D. (2013). *Color appearance models*. Hoboken, NJ: John Wiley & Sons, Ltd.
- Gabbard, J. L., Swan, J. E., Zedlitz, J., & Winchester, W. W. (2010). More than meets the eye: An engineering study to empirically examine the blending of real and virtual color spaces. In *Proceedings of the 2010 IEEE Virtual Reality Conference* (pp. 79–86). Washington, DC: IEEE Computer Society.
- Gelb, A. (1938). Colour constancy. In W. D. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 196–209). London, UK: Routledge & Kegan Paul.
- Gilchrist, A. (1977). Perceived lightness depends on perceived spatial arrangement. *Science*, 195(4274), 185–187.
- Gilchrist, A. (2006). *Seeing black and white*. Oxford, UK: Oxford University Press.
- He, Y., & Thorstenson, C. A. (2025). Facial color matching in optical see-through augmented reality. *Journal of Vision*, 25(10):16, 1–20, <https://doi.org/10.1167/jov.25.10.16>.
- Itoh, Y., Langlotz, T., Sutton, J., & Plopski, A. (2021). Towards indistinguishable augmented reality: A survey on optical see-through head-mounted displays. *ACM Computing Surveys*, 54(6), 1–36.
- Kingdom, F. A. (2011). Lightness, brightness and transparency: A quarter century of new ideas, captivating demonstrations and unrelenting controversy. *Vision Research*, 51(7), 652–673.
- Kleiner, M., Brainard, D. H., & Pelli, D. G. (2007). What's new in Psychtoolbox-3? *Perception*, 36 (Suppl.), 14.
- McCluney, R. (2014). *Introduction to radiometry and photometry* (2nd ed.). Boston, MA: Artech House, Inc.
- Mekni, M., & Lemieux, A. (2014). Augmented reality: Applications, challenges and future trends. *Applied Computational Science*, 20, 205–214.
- Metelli, F. (1970). An algebraic development of the theory of perceptual transparency. *Ergonomics*, 13(1), 59–66.
- Murdoch, M. J. (2020). Brightnessmatching in optical see-through augmented reality. *Journal of the Optical Society of America A*, 37(12), 1927–1936.
- Murdoch, M. J. (2025). From polaroid to augmented reality: The enduring advantages of white borders. *Electronic Imaging*, 37, 1–6.
- Murray, R. F. (2021). Lightness perception in complex scenes. *Annual Review of Vision Science*, 7, 417–436, <https://doi.org/10.1146/annurev-vision-093019-115159>.
- Peck, T. C., Good, J. J., Erickson, A., Bynum, I., & Bruder, G. (2022). Effects of transparency on perceived humanness: Implications for rendering skin tones using optical see-through displays. *IEEE Transactions on Visualization and Computer Graphics*, 28(5), 2179–2189.
- Schanda, J. (2007). *Colorimetry: Understanding the CIE system*. Hoboken, NJ: John Wiley & Sons.
- Schirillo, J., Reeves, A., & Arend, L. (1990). Perceived lightness, but not brightness, of achromatic surfaces depends on perceived depth information. *Perception & Psychophysics*, 48(1), 82–90.
- Shibata, T., Kim, J., Hoffman, D. M., & Banks, M. S. (2011). The zone of comfort: Predicting visual

discomfort with stereo displays. *Journal of Vision*, 11(8), 11, <https://doi.org/10.1167/11.8.11>.

Singh, M., & Anderson, B. L. (2002). Toward a perceptual theory of transparency. *Psychological Review*, 109(3), 492–519.

Zeiner, K., & Maertens, M. (2014). Linking luminance and lightness by global contrast normalization. *Journal of Vision*, 14(7):3, 1–15, <https://doi.org/10.1167/14.7.3>.

Appendix A: Creating the printed palette

We created the printed palette of Munsell ratings (Figure 3) as follows:

- Using custom software written in MATLAB, we generated a PostScript file of an image that was the same as Figure 3, except that it showed 11 squares with PostScript graylevels (achromatic colors) from 0.0 to 1.0 in steps of 0.1. We made a printout of this image (HP Color LaserJet Pro, model M452dw).
- With this printout at the same location and under the same illumination that the printed palette would have in the experiment, we measured the reflectance r of the 11 squares. We measured the luminance L of each square and the luminance L_{99} of a 99% reflectance standard (Spectralon SRS-99-020) at the same location as the square. We calculated the reflectance of each square as $r = 0.99L/L_{99}$.
- We found the least squares fit of a gamma function, $g = k(r - r_0)^\gamma$, to the 11 PostScript graylevels g and measured reflectances r , with free parameters k , r_0 , and γ .
- Using Equations 2 and 3, we converted the 14 Munsell values in the desired palette (i.e., 3.0, 3.5, 4.0, ..., 9.3; see Figure 3) to reflectances, and using the fitted function from step (c), we converted these reflectances to PostScript graylevels.
- Using custom MATLAB software, we generated a PostScript file of the palette shown in Figure 3, with the 14 PostScript graylevels calculated in step (d) assigned to the squares. We printed this image on the same printer used in step (a). This was the palette we used in the experiment.
- Using the same method as in step (b), we confirmed that the reflectances of the 14 squares on the palette printed in part (e) were correct. That is, we confirmed that the 14 squares had approximately the reflectances corresponding to

Munsell values 3.0, 3.5, 4.0, ..., 9.3, according to Equations 2 and 3. We found that the printed reflectances were within 0.015 reflectance units of the desired reflectances, except for the two lightest squares (labeled with Munsell values 9.0 and 9.3), where the printed reflectances were 0.053 below the desired reflectances. This means that the actual Munsell values of the 14 squares were within 0.1 of the labeled values, except for the two lightest squares, where the Munsell values were 0.25 lower than the labels. Thus, the printed reflectances were reasonably accurate. Furthermore, the two lightest squares, where the discrepancy was greatest, were used relatively seldom by participants; the square labeled 9.0 was chosen on 8% of trials, and 9.3 was chosen on 2%.

Appendix B: Converting Munsell value to reflectance

We used the approximation that achromatic Munsell value v and CIE lightness value L^* are related as $L^* = 10v$ (Fairchild, 2013, p. 82). The mapping from Munsell value v to reflectance r is then given by (Schanda, 2007, p. 61)

$$r = g\left(\frac{10v + 16}{116}\right) \quad (2)$$

where

$$g(x) = \begin{cases} \frac{108}{841} \left(x - \frac{4}{29}\right) & \text{if } x \leq \frac{6}{29} \\ x^3 & \text{otherwise} \end{cases} \quad (3)$$

To obtain this mapping from Schanda's formulas, we set $r = Y/Y_n$ and inverted the mapping from r to L^* .

Appendix C: Parameter b of the linear model

Suppose we have ordered pairs $\{(x_i, y_i)\}_{i=1}^n$ that we model as $\hat{y}_i = mx_i + b$. Least squares estimates of m and b are given by

$$\begin{aligned} \hat{m}, \hat{b} &= \operatorname{argmin}_{m,b} \sum_i (y_i - \hat{y}_i)^2 \quad (4) \\ &= \operatorname{argmin}_{m,b} \sum_i y_i^2 - 2x_i y_i m - 2y_i b + m^2 x_i^2 \\ &\quad + 2mbx_i + b^2 \quad (5) \end{aligned}$$

If $\sum_i x_i = 0$, the second-last term in Equation 5 disappears. Then there are no terms containing both m and b , so we can optimize the parameters independently. Eliminating terms containing m , the least squares estimate of b is given by

$$\hat{b} = \operatorname{argmin}_b \sum_i y_i^2 - 2y_i b + b^2 \quad (6)$$

$$= \operatorname{argmin}_b \sum_i (y_i - b)^2 \quad (7)$$

The value that minimizes this error is the sample mean, $\hat{b} = \frac{1}{n} \sum_i y_i$.

In the present experiment, the data consist of test patch luminances l_i and reflectance ratings r_i . The luminances l_i do not sum to zero. However, in Equation 1 we use normalized luminances $l'_i = l_i - l_0$, where l_0 is the mean of the luminances l_i . The normalized luminances l'_i do sum to zero, and so the fitted intercept parameter b equals the mean of the reflectance ratings r_i .