Title: Equivalent noise characterization of human lightness constancy

Authors: Vijay Singh1,2, Johannes Burge2,3,4,5, David H. Brainard2,3

1 Department of Physics, North Carolina A&T State University, Greensboro, NC, USA.

2 Computational Neuroscience Initiative, University of Pennsylvania, Philadelphia, PA, USA.

3 Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA.

4 Neuroscience Graduate Group, University of Pennsylvania, Philadelphia, PA, USA.

5 Bioengineering Graduate Group, University of Pennsylvania, Philadelphia, PA, USA.

**ABSTRACT:** A goal of visual perception is to provide stable representations of task-relevant scene properties (e.g. object reflectance) despite variation in task-irrelevant scene properties (e.g. illumination, reflectance of other nearby objects). To study such stability in the context of lightness, we introduce a threshold-based psychophysical paradigm. We measured how thresholds for discriminating the lightness of a target object (task-relevant property) are impacted by variation in the reflectance functions of background objects (task-irrelevant property). Our approach has roots in the equivalent noise paradigm which relates signal to noise properties of internal and external sources of noise and has been traditionally used to investigate contrast coding. For low background reflectance variation, the discrimination thresholds were nearly constant, indicating that in this regime observers’ internal noise determines threshold. As the background object reflectance variation increases, its effects start to dominate performance. We measured lightness discrimination thresholds as a function of the amount of variability in the background object reflectance function to determine the equivalent noise - the smallest level of task-irrelevant (i.e. background reflectance) variation that substantially corrupts the visual representation of the task-relevant variable (i.e. perceived object lightness). A computational model that uses a center-surround receptive field to estimate object lightness captures human behavior in this task. Our approach provides a method to characterize the effect of task-irrelevant scene variations on the perceptual representation of a task-relevant scene property.

**KEYWORDS:** Lightness, Equivalent Noise, Human Psychophysics, Color Vision

**INTRODUCTION:** To support effective action, vision needs to provide stable perceptual representations of the distal properties of objects. The processes that result in these representations start with the information provided by the proximal stimuli reaching the retinae. These proximal stimuli depend not only on the intrinsic properties of objects, but also on object-extrinsic properties of the visual scene, for example the illumination, the observer’s viewpoint, and the object’s position and pose. The challenge for the visual system is to recover stable correlates of object properties across variation in other scene variables. Understanding the degree to which the visual system does this, and how it does so, is an important goal of vision science.

Here we consider the perceptual task of representing the reflectance of an achromatic object embedded in a scene from the light reflected from the object and the rest of the scene to the eye. The perceptual correlate of the surface reflectance of achromatic objects is termed lightness. Computing a stable representation of object lightness poses a challenge to the visual system because the retinal irradiance of the image cast by the object varies both with the object’s overall reflectance and the irradiance of the illumination, as well as the position and pose of the object in the scene. The degree to which the visual system stabilizes the lightness representation against object-extrinsic variation determines the degree of lightness constancy achieved by the system.

The perceptual representation of lightness has been extensively studied using subjective psychophysical methods, in which observers report their subjective judgment of the object lightness (Arend & Goldstein, 1987; Arend & Spehar, 1993a, 1993b; Gilchrist, 1977; Toscani, Zdravkovic, & Gegenfurtner, 2016). At the broadest level, these results tell us that the visual system does indeed stabilize the perceptual lightness representation against changes in object-extrinsic factors in the scene (refs). There is less consensus, however, on how these perceptual computations should be understood (Adelson, 2000; Kingdom, 2011).

Although appearance methods have the advantage of tapping observers’ subjective experience rather directly, it can be challenging to relate data of this sort to underlying neural mechanisms (Stockman & Brainard, 2010)(ref). At the same time, there is mature theory that links objective measurements of psychophysical thresholds to the properties of physiologically measured neural responses (Parker & Newsome, 1998; Teller, 1984)(refs). To date, however, it has not been clear how to apply threshold measurements to questions of perceptual constancy. Indirect methods involve linking thresholds to appearance measurements, which have their origin in Fechner’s pioneering interpretation of Weber’s Law in terms of an underlying perceptual scale (Fechner, 1966; Hillis & Brainard, 2005, 2007; Nachmias & Sansbury, 1974)(ref; also refs). This approach holds promise, but there are documented cases where the threshold measurements fail to account for appearance effects related to lightness constancy (ref). Here we introduce an approach to using a psychophysical threshold paradigm to draw inferences about the psychophysical mechanisms underlying perceptual constancies. The approach has the feature that the results provide a behavioral basis for comparison to the precision of hypothesized neural representations of lightness.

Our approach begins with measurement of the ability of human observers to discriminate the lightness of two objects in the absence of any object-extrinsic variation. Next, we study how these discrimination thresholds change when object-extrinsic variation is introduced, which here takes the form of variability in the colors (i.e. reflectance spectra) of background objects in the scene (Brown & MacLeod, 1997; Lotto & Purves, 1999). The colors of background objects were randomized by sampling from a statistical model based on the reflectance spectra of natural surfaces. The amount of background color variation was controlled by a single model parameter. Finally, we measure discrimination thresholds as a function of the amount of background spectral variation. The discrimination thresholds quantify the difficulty of the lightness discrimination task. The change in thresholds from baseline (i.e. no background variation) quantifies the degree to which the object-extrinsic variation intrudes on the object-intrinsic representation.

As the variation in background color is increased, discrimination thresholds first remain nearly constant and then increase, with log squared threshold increasing linearly with log color variance. The minimum discrimination threshold and the color variance at which the threshold begins to rise are consistent across different observers. Moreover, a simple computational model that uses a center-surround receptive field to estimate the object lightness captures the essential features of human observers. The model allows us to quantify the effect of extrinsic variation on the observer’s representation of lightness, relative to the intrinsic precision of that variation.

**RESULTS:**

1. **Measurement of lightness discrimination thresholds**

We measured how variation in the reflectance spectra of background objects affect lightness discrimination thresholds using a two-alternative forced-choice (2AFC) design (Figure 1). On each trial, observers viewed a standard image and comparison image, presented on a calibrated monitor for 250ms each, one after the other with a 250ms inter-stimulus interval. (Figure 1a). The images were computer graphics renderings of 3D scenes; each scene contained an achromatic spherical target object. The observer’s task was to report the image in which the depicted target object was lighter. Across trials, we varied the luminous reflectance factor (LRF, (American Society for Testing and Materials, 2017 )) of the target object in the comparison image while keeping the LRF of the target object in the standard image fixed. The LRF is the ratio of the luminance of a surface under a reference illuminant (here CIE D65) to the luminance of the reference illuminant itself.

We recorded the proportion of times observers chose the comparison image as having the lighter target object at 11 values of the target object LRF. Figure 2 shows a psychometric function from a typical human observer. The proportion comparison chosen data was fit with a cumulative Gaussian using maximum likelihood methods (See Methods: Psychometric Function). The threshold was defined as the difference between the LRF of the target object at proportion comparison chosen 0.76 and 0.50, as determined from the cumulative Gaussian fit.

1. **Human lightness discrimination thresholds increase with background object reflectance variation**

To study the effect of background variation on lightness discrimination thresholds, we varied the reflectance spectra of the background objects in the images by sampling from a statistical model based on natural surface reflectance databases (See Methods: Reflectance and Illumination Spectra (Singh, Cottaris, Heasly, Brainard, & Burge, 2018)). Briefly, a database of natural surface reflectance functions (Kelly, Gibson, & Nickerson, 1943; Vrhel, Gershon, & Iwan, 1994) was projected along eigenvectors associated with the largest six eigenvalues of the dataset. These eigenvalues captured more than 90% of the variance in the database. The distribution of projection weights was approximated as a multivariate-normal distribution. Reflectance spectra of background objects were sampled from the multivariate-normal distribution and using the weights to construct spectra as the corresponding linear combination of the eigenvectors. The amount of variation in the background was controlled by multiplying the covariance matrix of the multivariate-normal distribution by a scalar.

We measured threshold as a function of the scalar that multiplied the covariance matrix. By varying the scalar from 0 (no variation) to 1 (variation in natural scenes), we can examine parametrically how background variation affects performance in the task. We generated images for six logarithmically spaced values of the covariance scalar. Figure 3 shows examples of images used in our psychophysical task for different choices of the covariance scalar. Discrimination thresholds were measured separately for each of the six values of the covariance scalar (Appendix: Table S2).

### Figure 4 shows how discrimination thresholds change with the amount of variability in the spectra of the background objects. We plot mean (across observers, N = 4) log threshold squared vs the log of the covariance scalar of the distribution. For low values of the covariance scalar, threshold is nearly constant. As the covariance scalar increases, log squared threshold rises approximately linearly with log covariance scalar, a dependence predicted by a simple model based on Signal Detection Theory (Figure 4; see below and Methods: Signal Detection Theory Model).

When the covariance scalar is 0, we conceptualize performance as limited by two factors (Pelli & Farell, 1999). One factor is the internal variability in the observer’s representation of target object lightness. The other factor is the efficiency with which the observer’s decision processes make use of the information provided by this representation. Our experiments cannot identify the relative contributions of these two conceptually distinct factors. In the following, we refer to both factors collectively as the observer’s internal noise for the lightness discrimination task.

As the covariance scalar increases, a third factor becomes important. This factor is the impact of external variability in background objects on observer’s representation of target object lightness. At low values of the covariance scalar, the internal noise dominates, and the impact of external variability has little effect on threshold. At high values of the covariance scalar, the impact of external variability limits performance, and thresholds increase systematically with increases in the covariance scalar. We interpret these effects further in the context of modeling introduced below.

Figure 5 shows the threshold variation for the individual observers. Each individual observer shows the same basic pattern as the mean across observers. Thresholds are constant for low values of the covariance scalar and, for higher values of the covariance scalar, thresholds rise approximately linearly on the log threshold squared versus log covariance plot. The most notable difference across individual observers is the slope of the rising limb of the measured functions.

1. **Quantifying impact of background surface variation on the lightness representation**

We modeled the psychophysical data with a framework based on the Signal Detection Theory (SDT, See Methods: Signal Detection Theory Model). In such models, performance is limited by two fundamental factors. The first factor is the response variability internal to the visual system (internal noise). The second factor is the effect of our experimentally induced stimulus variability of the background surfaces on the visual system’s representation of lightness (external noise). The models aid in estimating the effects of these two factors and evaluating how much external noise intrudes on performance, compared to the intrinsic precision of the visual system’s representation of target lightness. The model relates the discrimination threshold () with the variance in the internal noise (), the external noise ), and the covariance scalar () as:

where is the threshold with no external variation (see Methods: Signal Detection Theory Model for details). Intuitively, performance with no external variation (covariance scalar = 0.0) establishes the level of the internal noise, while the covariance scalar value corresponding to the threshold is double that with no external variation indicates when the level of the external noise is matches the level of internal noise.

To relate the SDT model directly to the stimuli used in our experiments, we developed a computational observer model based on a single-channel linear receptive field (see Methods: Computational Observer Linear Receptive Field (Lin-RF) Model). The Lin-RF model is an image-computable model that converts a high-dimensional input stimulus (i.e. the image) into a one-dimensional decision variable. The model calculates the response of a linear receptive field to an image to provide an estimate of the target LRV. The LRV estimate is used in the 2AFC paradigm to estimate the threshold of the computational observer, similar to the human psychophysics experiments. The receptive field model has the advantage that it can be implemented computationally and can be applied to stimuli with arbitrary noise properties.

Figure 4 shows the fit of the SDT model and its computational implementation (Lin-RF model) to the mean observer data. Figure 5 shows the model fits to the individual observer data. Both versions of the model capture the broad features of the data, although the computational implementation provides a better fit. This is because the computational implementation takes into account the fact that the actual covariance of the variation in background surface reflectances differs from the nominally specified variation, because we enforce a physical realizability constraint that surface reflectances lie between 0 and 1 (See Methods: Reflectance and Illumination Spectra).

By fitting the models to the human data by minimizing the mean squared error between the mean observed threshold and the model, we estimated the strength of internal and external variability of the human observers during this task. Figure 6 compares the standard deviation of internal and external noise for the analytical SDT model and the computational implementation. The estimates of standard deviation of the internal noise are consistent over the two versions and different observers (mean value of internal noise standard deviation = 0.0253, standard deviation = 0.0012 < 5% of mean, maximum deviation from mean = 0.0018 < 8% of mean). Both model versions show individual differences in the estimate of external noise among the observers. Across all observers, the estimate of external noise standard deviation is higher for the computational implementation observer model as compared to the TSD model.

**DISCUSSION:** Thelightnessof an object depends on the scene in which it lies. In this paper, our aim was to characterize the effect of color of background objects in a scene on object lightness. We used an equivalent noise approach for such characterization, an approach previously used to study the effect of noise in contrast detection. This approach allows one to relate the external noise power with the internal noise of the observer. We used this approach to develop a novel paradigm that relate the effects of variation in task-irrelevant scene properties on the perception of task-relevant properties. In our experiments, we measured human observers’ thresholds of discriminating two objects based on their lightness and determined how these thresholds change as the amount of variation in the color of background objects increases. The experimentally introduced variation was independent of the target object surface reflectance, the task relevant object property.

## Our results (Figure 4) show that when the variation in the color of background objects is small, the discrimination thresholds are nearly constant. In this regime, performance depends primarily on the internal noise of the observers’ decision-making process. As the amount of background color variation increases, the effect of external variation in the stimuli on observers’ representation of object lightness starts dominating and discrimination thresholds starts to increase. Using a computational model based on signal detection theory (SDT) framework, we related the thresholds to the variance of the internal and external variations associated with the task. Our model predicts that the amount of internal variation is consistent across observers, while the effect of external variation is observer dependent. The external variation makes the task difficult, leading to increase in discrimination thresholds, but the rate of increase in threshold is observer dependent. Figure 6 summarizes model predictions of the standard deviation in internal and external variations for human observers. Our work provides a systematic method to quantify the effect of variations in task-irrelevant properties on the perception of task-relevant property.

It is well known that a gray patch appears lighter against a uniform dark background as compared to the appearance of the same patch against a uniform white background. This failure of lightness constancy is generally associated with a center-surround structure of the visual receptive fields, which perform local comparison of the center with surround. As uniform color backgrounds are not common, visual receptive fields fail to estimate correct lightness. Our findings suggest that human observers find it hard to estimate object lightness even when the comparison is made against non-uniform backgrounds. The computational model captures the essential features of the human behavior and supports the center-surround receptive field idea.

Constancy (lightness/color ???) should improve in non-uniform background with multiple surfaces, as the brain can extract more information about the light source from the luminance from multiple surfaces. So, why do we observe an opposite effect? Perhaps the reason is that in our experiment we have fixed the light source. Thus, the luminance from the target object is linearly related to its reflectance. Background color variation corrupts the overall proximal stimulus, making lightness judgements difficult. The follow-up question would be to also include variation in light source to find out whether and by how much non-uniform backgrounds affect lightness perception in such conditions.

**CONCLUSIONS:** Our experiments show that variation in the color of background objects affects human observers’ perception of object lightness. Lightness estimates worsen as the amount of variation increases, as shown by the increase in lightness discrimination thresholds with increase in variance of background color variation. A computational linear receptive field model that estimates object lightness by projecting images over a center-surround receptive field and passing the response of the receptive field through a noisy decision-making process captures the essential features of observed human behavior at the discrimination task. The model relates the external scene variation to observers’ internal noise quantitatively.

**ACKNOWLEDGEMENTS**: NIH RO1-EY10016 (DHB), NIH R01-EY028571 (JB).

**METHODS:**

**Ethics statement.** All experimental procedures were approved by University of Pennsylvania Institutional Review Board and were in accordance with the World Medical Association Declaration of Helsinki.

**Preregistration**. The experimental design and the data analysis procedures for this study were preregistered before that start of the experiment. They are publicly available at: https://osf.io/7tgy8/. Deviations from and additions to the preregistered plan are described in the addendums to the pre-registration documents available at <https://osf.io/7tgy8/>.

The broad aim of the study was to study the effect of object extrinsic scene variations on human object lightness discrimination thresholds. For this, we pre-registered three experiments. The first experiment (pre-registered as Experiment 1) was abandoned because the task was too difficult. The findings of the second experiment (pre-registered as Experiment 2) provided control data and are reported in the Appendix. We focus in the paper on the third experiment (pre-registered as Experiment 3).

A notable deviation from the pre-registered plan for Experiment 2 was the change in the criteria to select observers for the experiment. The pre-registered criterion for selecting an observer for experiment 2 was that an observer “will be excluded if their mean threshold for the last two acquisitions run in the practice session exceed 0.025”. After, collecting data from 8 naive observers, we concluded that this criterion was too strict. Only one observer met the criterion. Hence, we modified the exclusion criteria as: “Observers will be excluded if their mean threshold for the last two acquisitions in the practice session exceeds 0.030.” The pre-registered plans also incorrectly mentioned that each image will be presented for 500ms instead of 250ms.

The pre-registration document also specifies the primary methods to analyze the data. It specified that the data would be analyzed separately for each observer by fitting a cumulative normal to the proportion comparison chosen data using the maximum likelihood method and that thresholds were to be extracted from the fit as the difference between object LRF at proportion comparison chosen 0.76 and 0.50. The observer thresholds at each level of background variability were to be measured three times and averaged. We indicated that the primary data feature of interest was the dependence of threshold on the covariance scalar, and predicted that thresholds would increase as the background variability increases. The modeling of the data, however, was developed post-hoc.

**Apparatus:** The stimuli were presented on a calibrated LCD color monitor (27-in. NEC MultiSync PA271W; NEC Display Solutions) in an otherwise dark room. The monitor was driven at a pixel resolution of 1920 x 1080, a refresh rate of 60Hz, and with 8-bit resolution for each RGB channel. The host computer was an Apple Macintosh with an Intel Core i7 processor. The experimental programs were written in MATLAB (MathWorks; Natick, MA) and relied on routines from the Psychophysics Toolbox (<http://psychtoolbox.org>) and mgl (<http://justingardner.net/doku.php/mgl/overview>). Responses were collected using a Logitech F310 gamepad controller.

The observer’s head position was stabilized using a chin cup and forehead rest (Headspot, UHCOTech, Houston, TX). The observer's eyes were centered horizontally and vertically with respect to the display. The distance from observer's eyes to the monitor was 75cm.

**Monitor Calibration:** The monitor was calibrated using a spectroradiometer (PhotoResearch PR650). To calibrate the monitor, we focused the spectroradiometer on a patch displayed on the center of the monitor. The patch size was 4.8cm x 4.6cm (3.67° x 3.51°). The optics of the radiometer sampled the emitted light from a 1° circular spot within the patch. The spectral power distribution of the three monitor primaries was measured in the range 380nm to 780nm at 4nm steps. The spectral power distribution was measured for 26 values of the input applied to the primaries. The applied values were in the range [0, 1] spaced 0.04 apart, where 1 corresponds to the maximum value of the allowed input and 0 corresponds to no input. The power distribution of the three primaries were also measured at 32 different combinations of the input in the range [0,0,0] to [1,1,1]. These measurements were compared to the applied input settings to check the linearity of the primaries. The maximum absolute deviation of the x-y chromaticity between the applied and measured values were less than 0.0028 and 0.0027 for x and y chromaticity respectively, and less than 1% for luminance.

**Observer Recruitment and Exclusion:** Observers were recruited from the University of Pennsylvania and the local Philadelphia community and were compensated for their time. Observers were screened to have normal visual acuity (20/40 or better) and normal color vision, as assessed with pseudo-isochromatic plates (S, 1977). These exclusion criteria were specified in the pre-registration document. One observer was discontinued at this point as they did not meet the normal visual acuity criterion.

Observers who passed the vision screening then participated in a practice session. This session also served to were screen for observers’ ability to reliably perform the psychophysical task. This screening was performed in the first session for each observer, which was considered a practice session. At the beginning of the practice session, observers were familiarized with the task. For this they performed a familiarization acquisition (See Methods: Experimental Detailsfor the definition of an acquisition). In the familiarization acquisition, observers performed 40 trials of the task using images with covariance scale factor 0.00 (10 easy trials, 10 moderate trials, and 20 regular trials). In the easy trials, the observers compared images with target object luminous reflectance factor (LRF) 0.35 and 0.45. In the moderate trials, they compared images with target object LRF 0.40 to images with target object LRF 0.35 or 0.45. In the regular trials they compared images with target object LRF 0.40 to images with target object LRF in the range [0.35, 0.45]. The data from the familiarization acquisition was not saved. After this the observer performed three normal acquisitions for images with covariance scale factor 0.00. At the end of the practice session, the mean threshold of the observer for the last two acquisitions was computed. The observer was excluded from further participation if their mean threshold for the last two acquisitions in the practice session exceeded 0.025 (log T2, -3.2). This exclusion criterion was specified in our pre-registered protocol.

Observers who met the performance criterion participated in the rest of the experiment. Observers performed only one session on a given day. The sessions were scheduled as per the availability of the experimenter and the observer. The data of all observers in the main experiment (pre-registered Experiment 3) was collected over a period of 4 weeks.

A total of 17 observers participated in the practice sessions for Experiments 2 and 3. To de-identify observer information in the data, observers were numbered in the order they performed the practice sessions. 10 observers participated in the practice sessions for Experiment 3 (6 Female, 4 Male; age 18-56; mean age 30.7). Four of these observers (Observer 2, Observer 4, Observer 8 and Observer 17) met the performance criterion set for screening (2 Female, 2 Male; age 23-56; mean age 38.25). All observers had normal or corrected-to-normal vision (20/40 or better in both eyes, assessed using Snellen chart) and normal color vision (0 Ishihara plates read incorrectly). Observers were dark adapted before performing the experiments. The choice of four observers to complete the experiment was specified in our pre-registered protocol.

**Stimulus Design:** We measured lightness discrimination thresholds as a function of the amount of variability in the surface reflectance of the background objects. The reflectances were chosen from a distribution of natural surfaces. The amount of variability was controlled by multiplying the covariance matrix of the distribution by a scalar (See Methods: Reflectance and Illumination Spectra). We measured thresholds for six logarithmically spaced values of the covariance scalar.

For each value of the covariance scalar, we generated a dataset of 1100 images. The dataset had 100 images each at 11 values of the target object LRF. The LRF of the target object in the standard images was 0.4 and the lightness in the comparison image varied between 0.35 and 0.45 at steps of 0.01 (11 comparison levels). We generated 100 images at each comparison level, each with a different choice of the reflectance spectra of the background scene objects. For scale factor 0.00 we generated a set of 11 images, one at each LRF level, as the background remained fixed in this case. All images were generated without secondary reflections specified in the rendering process. The spectral power distribution of each light source in the scene was fixed over all images. We chose this to be the standard daylight spectrum D65 (See Methods: Reflectance and Illumination Spectra). The geometry of the 3D scene was also held fixed.

**Experimental Details:** We used a two-interval forced choice procedure to measure thresholds. We showed two images, one after the other, on a calibrated computer monitor and asked the observer to report the interval in which the target object was lighter. We fixed the reflectance of the target object in the standard image and varied the reflectance of the target object in the other image, which we refer to as the comparison image. The method of constant stimuli was used. The temporal order in which the standard and comparison images were presented was randomized on each trial. An audio feedback was provided after every trial.

We define a trial as the presentation of the two images (standard and comparison images) and collection of the observer’s response. We define an interval as the presentation of one of the images in the trial. Thus, a trial has two intervals.

The experiment was structured as follows. We define an acquisition as the data collected at one covariance scale factor with 30 trials at each of the 11 comparison levels. We define a permutation as a set of six acquisitions, where each acquisition corresponds to one of the possible six scale factors. We collected three permutations for each observer, with a new random order drawn for each permutation. Thus, after the practice session (see: Recruitment and Exclusion), there were total 18 acquisitions. We divided these 18 acquisitions over 6 sessions, each session with 3 acquisitions. In each acquisition, we randomly selected the images on the trials from the pre-generated image databases. The first five trials of each acquisition were moderate trials (as defined above in Observer Recruitment and Exclusion) to acclimatize the observer to the experimental task. The responses for these five trials were not saved.

Each acquisition thus consisted of 330 trials (excluding the 5 moderate acclimatization trials), 30 at each of the 11 comparison levels. The trial sequence (order of comparison stimuli) in an acquisition was generated pseudo-randomly at the beginning of the acquisition. For this, at each comparison lightness level, 30 standard and comparison images were chosen pseudo-randomly with replacement from the image dataset. The sequence of presentation of these 330 trials were randomized and saved. For each trial, the order of presentation of the standard and comparison image was also determined pseudo-randomly and saved. The trials were presented according to the saved sequence.

The trials in an acquisition were presented in three blocks of 110 trials each. At the end of each block observer took a rest (of minimum 1 minute). The observer could terminate the experiment anytime during the acquisition. If an observer terminated an acquisition, the data for that acquisition was not saved. No observer terminated any acquisition. One observer rescheduled at the beginning of a session due to tiredness for reasons unrelated to the experiment. The session was rescheduled.

At the beginning of the first experimental session (the practice session) for an observer, the experimenter explained the experimental procedures and obtained consent for the experiments. The experimenter then tested the observers for normal visual acuity and color vision. The observers were then taken to the dark room where the observers were described the task and familiarized with the display, chin rest, and response box. Once familiar, the observers were dark adapted (by sitting in the dark room for approximately 5 minutes). Once ready, the observers performed the familiarization acquisition. After the familiarization acquisition, the observers performed the other three acquisitions of the practice session. The entire practice session took nearly one hour.

The observers who met the criteria performed 18 acquisitions over 6 other sessions. The order of these acquisitions was determined pseudo-randomly at the beginning of the practice session. In each session, the observer performed only three acquisitions. The observers were dark adapted at the beginning of each session.

**Stimulus Presentation:** The size of each image was 2.6cm x 2.6cm on the monitor, corresponding to 2° by 2° visual angle. The target object size on the screen in the 2D images was ~1° in diameter. Each image was presented for 250ms (this was a deviation from the preregistration document, which specifies the presentation time as 500ms), with an inter-stimulus interval of 250ms and inter-trial interval of 250ms. Inter-stimulus interval (ISI) is defined as the interval between the first and the second image presented on each trial. The response for each trial was collected after both the images had been displayed and removed from the screen. The observer could take as long as they wished before entering the response. Feedback was provided via tones presented after the response. The next trial was presented 250ms (ITI) after the feedback. Thus, the actual inter-trial interval depended on the response time of the observer.

**Image Generation:** The images were generated using software we refer to as Virtual World Color Constancy (VWCC) (<https://github.com/BrainardLab/VirtualWorldColorConstancy>). VWCC is written using MATLAB. It harnesses the Mitsuba renderer to render simulated images from scene descriptions, and also takes advantage of our RenderToolbox package (rendertoolbox.org; (Heasly, Cottaris, Lichtman, Xiao, & Brainard, 2014)). To render an image, we first create a 3D model that specifies the base scene. Objects and light sources can be inserted in the base scene at user specified locations. The 3D models were based on a base scene provided as part of RenderToolbox and modified using Blender, an open-source 3-D modeling and animation package (<https://www.blender.org/>). Next, we assigned reflectance spectra and spectral power distribution functions to the objects and light sources in the scene (see Reflectance and Illumination Spectra Generation for how these spectra were generated). Once the geometrical and spectral features were specified, we render a 2D multispectral image of the scene using Mitsuba, a physically-realistic open-source rendering system ([https://www.mitsuba-renderer.org](https://www.mitsuba-renderer.org/), (Jakob, 2010)). The images were rendered at 31 wavelengths equally spaced between 400nm and 700nm. The images were rendered with the camera field of view of 17° with an image resolution of 320-pixel by 240-pixels with the target object at the center. A 201-pixel by 201-pixel area, centered around the spherical target object, was cropped for display on the monitor.

To present the multispectral images on the monitor, they were first converted to LMS images using the Stockman-Sharpe 2° cone fundamentals (T\_cones\_ss2 in the Psychophysics Toolbox). Then the monitor calibration data and standard methods (Brainard, 1989) were used to convert the LMS images to RGB images. Finally, a common scaling was applied to all of the images to bring them into the display gamut of the monitor. The gamma corrected RGB images was presented on the monitor during the experiment.

**Reflectance and Illumination Spectra:** The reflectance spectra for the objects were generated using random sampling of datasets of natural world objects as described in Singh et. al ((Singh et al., 2018)). We first approximated the natural datasets using principal component analysis (PCA). We projected the dataset along the PCA eigenvectors with the largest 6 eigenvalues. For the reflectance spectrum dataset, these directions capture more than 90% of the variance. We then approximated the resulting distribution by a multi-normal distribution. Reflectance spectra for the objects in the scene were generated using random sampling from this multi-normal distribution. The reflectance spectra were constructed as a linear combination of PCA eigenvectors and the sampled weights. The amount of variation in the color of the background objects was controlled by multiplying the covariance matrix of the distribution with a scalar. We generated images for six logarithmically spaced values of the covariance scalar [0, 0.01, 0.03, 0.1, 0.3, 1.0]. We imposed a physical realizability condition on the spectral samples by ensuring that the reflectance at each spectral frequency was within 0 and 1. Due to this condition, the variance of the generated spectral samples for some covariance scalars was lower than the variance of the multi-normal distribution.

The power spectrum of the light sources was chosen as standard daylight D65 spectrum. We normalized the D65 spectrum by its mean power to get the relative spectral shape. This spectral shape was scaled by a fixed scalar to get the power spectrum. The same relative spectral shape and scalar was applied to the power spectrum of all light sources in the visual scene. To find the scalar we first rendered the entire image database by setting it equal to 5 (chosen arbitrarily to render images using Virtual World Color Constancy pipeline). Then, we used monitor calibration data and the image dataset to find out the maximum pixel value that needed to be displayed on the monitor during the experiments. We chose the scalar such that the maximum value to be displayed would be 0.9 of the range allowed by the monitor gamut. The entire image dataset was scaled by the same scalar.

**Psychometric Function:** The proportion comparison chosen data was used to obtain the psychometric function for each acquisition. Each acquisition consisted of 330 trials with 30 trials at each comparison lightness level. At each lightness level, we recorded the number of times the observers chose the comparison image to be lighter. The proportion comparison chosen data was fit with a cumulative Gaussian using the Palamedes toolbox (Prins & Kingdom, 2018). The data was fit to obtain all four parameters of the psychometric function: threshold, slope, lapse rate and guess rate. While estimating the parameters, the lapse rate was set equal to the guess rate and was forced to be in the range [0 0.05]. The model was fit to the data using maximum likelihood method. The threshold was obtained as the difference between the LRFs at proportion comparison chosen 0.76 and 0.50 as obtained from the cumulative gaussian fit.

**Signal Detection Theory Model:** We developed a model of performance in our task, based on the signal detection theory (Green, 1996). We model the visual response to each image by a univariate internal representation denoted by the variable . This variable depends on the image and is perturbed by noise. We assume that for any fixed image, is a Gaussian distributed random variable whose mean depends on the target object LRF. For each image, we assume that is perturbed on a trial-by-trial basis by independent zero-mean Gaussian noise, and we assume that the variance this noise is the same for the response to all images. We refer to the noise that perturbs for a fixed image as the internal noise and denote its variance as For each trial of the experiment, takes on two values, and , one for the interval containing the standard and the other for the interval containing the comparison.

If we consider performance for a particular pair of target standard and comparison LRFs, performance depends both on the difference between the expected values of for each pair of LRFs, and respectively and on the value of In our experimental design we have an ensemble of images corresponding to each value of the target sphere LRF. The fact that we draw stochastically from this ensemble on each trial introduces additional variability into the value of the decision variable that corresponds to a fixed target LRF. We call this the external variability, and model it as a Gaussian random variable with zero mean and variance We assume that depends on the experimentally chosen covariance scalar, but not on the target sphere LRF. Thus, the distributions of and , for a particular choice of target standard and comparison LRF and covariance scalar, are given by and . Here is the mean value of the internal representation to the standard image and is the mean value of the internal representation to the comparison image. is the total variance, given as , where and are the variance of the internal and external noise, respectively.

In the standard formulation of Signal Detection Theory for a 2AFC task, the observer makes their decision based on a comparison of and , choosing the interval with the higher value of . The observer’s sensitivity depends on the mean values and the variance and is captured by the quantity d-prime given as . This quantity () measures the distance between the two distributions by number of standard deviations units. indicates the inability to distinguish between the standard and the comparison image. Larger values of indicates better ability to distinguish between the standard and the comparison image.

For a fixed value of , the difference in mean values is directly proportional to the standard deviation :

We further assume that the difference in mean value of the internal variable is proportional to the difference in the LRFs of the target object in the standard and comparison images (). That is, , where is the proportionality constant. Then we have,

When we measure threshold in a 2AFC task, we choose a criterion proportional correct and find the that corresponds to that proportion correct. Our choice of 0.76 corresponds to In addition we can choose , in essence setting the units for to be those of the experimentally determined target LRF.

In our experiment, the external variability was induced by changing the surface reflectance of the objects in the background. We used a multinormal distribution to generate surface reflectance function of the background objects. To change the amount of external noise, we scaled the variance of the multinormal distribution by multiplying the covariance matrix with a scalar. Thus, to use the above relationship for the data collected in our experiments, we need to modify it as follows:

where is the covariance scalar and is the external noise introduced when the ensemble of images for each value of target LRF has the reflectance of the background surfaces drawn from our model of natural surface reflectances.

Converting the equation above to the form we use to represent the data, we have

The equation above predicts that form of the threshold as a function of covariance scalar should increase monotonically. For small values of (/), the threshold will approach a constant giving For large values of (/), the quantity will approach a straight line of slope 1 in the plot. Fitting such measurements allows us to check whether the model describes the data as well as to determine the two parameters and . In particular, we can establish the relative contribution of the internal representational variability and external stimulus drive variability in limiting lightness discrimination. Indeed, the parameter quantifies how much the variation in background surface reflectance intrudes on the internal representation that mediates the lightness discrimination task, in a manner that may be compared to the intrinsic precision of that representation specified by

**Computational Observer Linear Receptive Field Model:** When the external noise added to the images is characterized by a multivariate Gaussian, a simple linear receptive field model of the visual system is equivalent to the SDT model developed above. We first develop this equivalence. The advantage of the receptive field formulation is that it can be implemented computationally and applied in cases where the external noise is not Gaussian. In our case, the fact that we truncate surface reflectances to lie between 0 and 1 to satisfy physical realizability means that the Gaussian characterization is only an approximation, so that adopting the linear receptive field formulation improves the precision of our modeling. This approach also allows us to incorporate the Poisson variability of the cone excitations.

Denote the stimulus image by the column vector , and the receptive field by the column vector . The entries of are the radiant power emitted by the monitor at each image location. The entries of are the corresponding sensitivities of the linear receptive field to each entry of . The response of the receptive field is given as , where is a random variable representing a draw of the internal noise (variance ) of the visual system.

Denote and as the standard and comparison images without external noise. External Gaussian noise is added to both and , with covariance matrix . The external noise need not have zero mean. After incorporation of the external noise, the response of the receptive field to the comparison and standard images is given by

Here is a random variable representing a draw of external noise in the image space, while represents the internal noise, and is a random variable representing the overall effect of the external and internal noise. Since the receptive field and noise models are linear and Gaussian, is Gaussian with variance

The mean difference between the receptive field response to the comparison and the standard image is given by . Here and are the standard and comparison images without external noise added and is a constant.

We associate the linear receptive field response with the internal representation of the SDT model developed above. That is, we assume that on each trial, the observer chooses as lighter the interval for which the response of the receptive field is greater. Following the development in the SDT model section above, we have

where we have introduced the covariance scalar in the term corresponding to the variance in the external noise and where denotes the covariance matrix of the external noise corresponding to the level of variation in natural images. Comparing to relation derived in the section Theory of Signal Detection, we see that this is the same functional form for the relation between and as derived there, where we associate and .

To fit this model, we use a one-parameter description of a simple center-surround receptive field and use simulation to compute model responses for any choice of . Once the fitting procedure (described below) establishes and that best account for the data, we then find directly by passing the images corresponding to through the receptive field and finding the resulting variance.

**SDT Model Fit:** Thetheory of signal detectionmodelwasfit to the threshold versus covariance scalar data to obtain the parameters and . The parameters were obtained by minimizing the mean squared error between the measured and predicted threshold using the MATLAB built-in function *fmincon*. (MATLAB scripts are provided as supplementary documents.)

**Linear Receptive Field Model Fit:** We fit the linear receptive field model using a simulation approach, so that we can incorporate a model of the early visual system into the computations, and so that we handle accurately the truncation of the Gaussian model of natural surface reflectances.

The model of early visual system to the image database was estimated as described in (Singh et al., 2018). The model incorporated typical optical blurring, axial chromatic aberration (Marimont & Wandell, 1994), and spatial sampling of the long (L), middle (M) and short (S) wavelength- sensitive cones (Brainard, 2015). The L:M:S cone ratio in the cone mosaic was chosen to be 0.6:0.3:0.1 (1523 L cones, 801 M cones, 277 S cones). The CIE physiological standard [ref] was used to get the spectral sensitivities of the cones. The response of the cones was calculated as the number of photopigment isomerizations in 100ms, including the Poisson nature of the isomerization (Hecht, Shlaer, & Pirenne, 1942). The model was implemented using the software infrastructure provided by ISETBio (https://isetbio.org). The cone responses were demosaiced using linear interpolation to get the response of each cone class over the entire image. Further, the response of each cone class was normalized by the summed (over wavelength) quantal efficiency of the corresponding cone class to make the magnitude of the three cone classes similar to each other.

The dot product of the cone response image was taken with a center-surround linear receptive field. The receptive field was square in shape and its size was equal to the size of cone response images. The center of the receptive was a circle of radius the size of the image of the target object. The central region was taken to have spatially uniform positive sensitivity, while the surround was taken to have spatially uniform negative sensitivity. Each point in the central region had sensitivity and each region of the surround had sensitivity , which was a parameter of the model. The RF was copied three times; one each for the L, M and S cone response images. The mean RF response was estimated as the sum total of the dot product of the RF with the L, M and S cone response images. Gaussian noise (representing noise in the decision-making process) with zero mean was added to the resulting dot product. The variance of the decision noise () and the value of the RF surround () were the two parameters of the model.

The threshold of the computational model was obtained using a two-interval force choice paradigm similar to the experiment. For each trial, we sampled a standard image and a comparison image from our dataset at random. We obtained the response of the receptive field (noise-added dot product) to the images and compared them to predict the image with lighter target object. This process was repeated 10,000 times for each of the 11 comparison LRF level. The proportion comparison chosen data was used to get the psychometric function and the threshold of discrimination, similar to the method used for human observer data. We estimated the threshold at the six values of covariance scalar at which we performed the human experiments. We calculated the mean squared error (averaged over the six covariance scalar values) between the thresholds of the human observer and the computational model for a large set of values of the two model parameters: the variance of the decision noise () and the value of the RF surround (). The mean squared error values obtained as a function of these two parameters were fit with a degree two polynomial of two variables using the MATLAB *fit* function. The resulting polynomial was minimized to get the parameters with lowest mean square error. These parameters were used to estimate the internal and external noise standard deviation for the computational observer using the relations: and as explained above, where the constant was obtained using the relationship . The best parameters and the internal and external noise standard deviation were estimated separately for the mean observer and the individual observers.

The retinal images and the MATLAB function to get the model thresholds are provided as supplementary documents.

**Code and Data Availability:** Observers response in the psychophysics task and their thresholds are provided in the supplementary documents. The SI also provides the MATLAB scripts to generate Figures 2, 4, 5 and 6 and the scripts to get model thresholds. The retinal images are provided as .mat files in a zip folder.

**Figure 1:** **Psychophysical task.** **(a)** On every trial of the experiment, human observers viewed two images in sequence, a standard image and a comparison image and indicated the one in which the spherical target object in the center of the image was lighter. Example standard and comparison images are shown. The images were computer graphics simulations. The simulated reflectance functions of the target were spectrally flat and the spheres appeared gray. The overall reflectance of the target was held fixed in the standard images, and differed between standard and comparison. Performance (proportion correct) was measured as a function of this difference to determine discrimination threshold. The reflectance functions of objects in the background could be held fixed or vary between standard and comparison on each trial (as illustrated here). The order of presentation of the standard and comparison images was randomized from trial to trial. Discrimination thresholds were measured as function of the amount of variation in background object reflectance.

**(b)** Trial sequence. RN-1 indicates the time of the observer’s response for the (N-1)th trial. The Nth trial begins 250ms after that response (Inter Trial Interval, ITI). The Nth trial consists of two 250ms stimulus presentation intervals with a 250ms inter-stimulus interval (ISI). The observer responds by pressing a button on a gamepad after the second stimulus has been shown. The observer can take as long as he or she wishes before making the response, with an example response time denoted by RN in the figure. The next trial begins 250ms after the response.

**Figure 2: Psychometric function.** We recorded the proportion of times the observer chose the target in the comparison image to be lighter, as a function of the comparison LRF. The LRF of the target object in the standard image was fixed at 0.4. The LRF of the target object in the comparison image were chosen from 11 linearly spaced values in the range [0.35, 0.45]. Thirty trials were presented at each comparison LRF value. We fit a cumulative normal distribution to the proportion comparison chosen data using maximum likelihood methods. The guess and lapse rates were assumed to be equal and were restricted to be in the range [0, 0.05]. The threshold was measured as the difference between the LRF at proportion comparison chosen equal to 0.7604 and 0.5, as predicted by the cumulative normal fit. This figure shows the data for Observer 2 for scale factor 0.00 in the first experimental session for that observer. The point of subjective equality (PSE, the LRF corresponding to proportion chosen 0.5) was close to 0.4 as expected and the threshold was 0.0233. The lapse rate for this fit was 0.05.

**Figure 3: Variation in background color:** The reflectance spectra of background objects were chosen from a multivariate Gaussian distribution that modeled the statistics of natural surface spectra. The variation in the reflectance spectra was controlled by multiplying the covariance matrix of the distribution with a scalar. We generated images at six levels of the scalar. Each column shows three sample images at each of the six values of the scalar. The leftmost column corresponds to no variation and the rightmost column corresponds to the modeled variation of natural surfaces. The target object (sphere at the center of each panel) in each image has the same LRF. For each value of the scalar, we generated 1100 images, 100 each at 11 linearly spaced target LRF levels across the range [0.35, 0.45]. Discrimination thresholds were measured separately for each value of the covariance scalar shown.

**Figure 4: Background variation increases lightness discrimination threshold.** Mean (N = 4)log squared threshold vs log covariance scalar from the human psychophysics (red circles). The error bars represent +/- 1 SEM taken between observers. The data were fit with the function (SDT Model) with (red curve). The best fit parameters are indicated in the legend. The threshold of the linear receptive field (Lin-RF) model was estimated at 10 logarithmically spaced values of the covariance scalar (black squares). The black smooth curve is an approximation to these points of the functional form where and , , and are parameters.

**Figure 5: Threshold of individual human observers.** Mean (across sessions) squared threshold vs log covariance scalar for individual human observers. Same format as Figure 4; here the error bars represent +/- 1 SEM taken across sessions for each observer. The parameters of the SDT model and the Lin-RF models were obtained separately for each observer.

**Figure 6: Internal and external noise standard deviation for human observers.** Noise standard deviation for human observers estimated using SDT model and the computational linear receptive model (Lin-RF) model. While the internal noise estimates are consistent over the two models, the external noise estimated by the Lin-RF model is higher compared to the SDT model.

**APPENDIX:**

**Measurement of human object lightness discrimination thresholds under variation in object background:** This supplemental experiment, pre-registered as Experiment 2, provided preliminary data that helped shape the design of the main experiment presented in the paper (which was Experiment 3 of the pre-registration documents). It aimed to determine whether variation in the reflectance of background objects had an effect on human lightness discrimination thresholds. It established that human object lightness discrimination thresholds increase if the reflectance of background objects vary, as compared the case when the discrimination is made against the same background. It also studied the effect of inclusion or not of secondary reflections in the rendering process as well and assessed the effect of implementing background variation across trials rather than across intervals.

The basic methods were the same as the experiment described in the main paper. The practice session was conducted with the images in Condition 1 described below. The observers were retained for the experiment if their average threshold of the last two acquisitions during the practice session was lower than 0.030. This was a deviation from the pre-registered plan where we set the threshold criterion as 0.025. After collecting data from 8 observers, we realized that the criterion was too strict. Only one observer had met the criterion. After modifying the threshold criterion, we included two of the initially discontinued observers in our experiment (Observer 5 and Observer 8). Total of 11 naïve observers participated in the practice sessions. Four of these observers met the criteria for continuing the experiment. Two of these observers also participated in the main experiment (Observer 4 and Observer 8).

We measured lightness discrimination threshold of four naïve human observers using a two-interval forced choice paradigm. The thresholds were measured for three specific types of background variation (Figure S1). The reflectance spectra of the background objects were generated with the covariance scale factor set to 1. These three conditions were:

*Condition 1.* *Fixed background:* In this condition, the spectra of objects in the background were kept fixed for all trials and for all intervals. We generated 11 images, one at each comparison LRF level.

*Condition 2. Between-trial background variation*: In this condition, the spectra of the objects in the background were the same for the two intervals within a trial but varied from trial-to-trial.

*Condition 3. Within-trial background variation*: In this condition, the spectra of the objects in the background varied between trials as well as between the two intervals of a trial. The background variation corresponded to covariance scale factor equal to 1.

In Conditions 2 and 3, the light reflected from the target object varied from image to image (even at the same LRF level of the target object) because of secondary reflection of light coming from the background objects was included in the rendering. We also measured the thresholds without secondary reflections for these two conditions. We call these conditions Condition 2a and 3a.

*Condition 2a. Between-trial background variation without secondary reflection*: Same as Condition 2, but without multiple reflections of light from object surfaces. The light rays only bounce off once from the surfaces before coming to the camera.

*Condition 3a. Within-trial background variation without secondary reflections*: Same as Condition 3, but without multiple reflections of light from object surfaces. Condition 3a was the same as the experiment reported in the main paper for covariance scalar equal to 1.

**Results:** Figure S2 shows the discrimination thresholds of the four human observers for the five conditions studied in this experiment. We plot the mean threshold and the standard error of the mean (SEM) taken over the three separate threshold measurements. For each observer, the thresholds for Condition 3 and 3a were higher compared to Condition 1, 2 and 2a. The average increases in threshold of the observers for Conditions 3 and 3a as compared to Condition 1 (baseline) were 79% and 60% respectively. The average increases in threshold for Conditions 2 and 2a were much smaller, 13% and 17% respectively. The thresholds for Condition 1, 2 and 2a were nearly within one SEM of each other (averaged over the observers and three conditions). On the other hand, the thresholds for Conditions 3 and 3a were respectively (on average) 7.2 and 5.4 SEM larger than the threshold of Condition 1. The thresholds without secondary reflections (Conditions 2a and 3a) were within one SEM from the conditions with secondary reflections (Conditions 2 and 3). This preliminary experiment established that lightness discrimination thresholds are higher for the case when the two objects are being discriminated against different backgrounds compared on the same trial, as compared to when the backgrounds are the same within trial. Trial-to-trial variability in background across trials has little, if any, effect. The effect is similar when the rendering is performed with and without secondary reflections, indicating the effect is due to the spectral change in the background and not due to the variation in the amount of light being reflected from the target object surface. In the main experiments, we rendered without secondary reflections to avoid introducing such variability. Figure S2 also shows the threshold of the observers in Experiment 3 for the condition with covariance scalar equal to 1. This condition is equivalent to Condition 3a of Experiment 2. The thresholds of the observers were consistent across the two measurements.

Table S1: Observer Thresholds for Experiment 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Mean Threshold +- SEM (averaged over sessions) | | | | |
| Observer | Condition 1 | Condition 2 | Condition 2a | Condition 3 | Condition 3a |
| 4 | 0.0269+-0.0013 | 0.0254+-0.0013 | 0.0235+-0.0011 | 0.0366+-0.0030 | 0.0330+-0.0018 |
| 5 | 0.0217+-0.0005 | 0.0305+-0.0039 | 0.0300+-0.0017 | 0.0382+-0.0031 | 0.0389+-0.0022 |
| 8 | 0.0167+-0.0011 | 0.0169+-0.0020 | 0.0175+-0.0017 | 0.0325+-0.0016 | 0.0273+-0.0016 |
| 11 | 0.0252+-0.0013 | 0.0268+-0.0018 | 0.0285+-0.0002 | 0.0525+-0.0038 | 0.0439+-0.0068 |

**Table S2. Lightness discrimination thresholds for Experiment 3**: Mean threshold (averaged over sessions) standard error of measurement of four human observers measured at six logarithmically spaced values of covariance scalar.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Observer | Covariance Scalar | | | | | |
| 0 | 0.01 | 0.03 | 0.1 | 0.3 | 1 |
| 2 | 0.0217+-0.0009 | 0.0238+-0.0006 | 0.0307+-0.0036 | 0.0294+-0.0008 | 0.0392+-0.0005 | 0.0429+-0.0049 |
| 4 | 0.0241+-0.0035 | 0.0215+-0.0015 | 0.0271+-0.0019 | 0.0246+-0.0018 | 0.0299+-0.0020 | 0.0295+-0.0014 |
| 8 | 0.0266+-0.0019 | 0.0214+-0.0005 | 0.0221+-0.0008 | 0.0273+-0.0024 | 0.0269+-0.0020 | 0.0318+-0.0041 |
| 17 | 0.0224+-0.0020 | 0.0236+-0.0030 | 0.0315+-0.0024 | 0.0347+-0.0027 | 0.0390+-0.0046 | 0.0454+-0.0032 |

**Figure S1:** Example stimuli for Conditions 1, 2 and 3 in Experiment 2 to study the effect of background color on lightness discrimination threshold. In condition 1, the background was fixed in every trail and every interval. In condition 2, the background varied from trial to trial, but remained fixed in the two intervals of a trial. In condition 3, the background varied in each trial and interval. For illustration, in this figure we have chosen the stimulus on the left to be the standard image with target object at 0.4 LRF and the on the right to be comparison image with target object at 0.45 LRF. In the experiment, the two images were presented sequentially in random order at the center of the screen. Conditions 2a and 3a stimuli are similar to condition 2 and 3 respectively, but without secondary reflections.

**Figure S2:** Lightness discrimination threshold of four human observers in the five conditions in Experiment 2 (The data points have been jittered to avoid marker overlaps). The thresholds are higher for the condition where the objects are compared against different backgrounds (Condition 3 and 3a) as compared to the same background (Condition 1, 2, 2a). Secondary reflections do not have any significant effect on thresholds (Condition 2a and 3a). Condition 3a of Experiment 2 is equivalent to the condition with covariance scalar equal to 1 (). The thresholds for this condition are also provided for comparison. Two observers from Experiment 2 also participated in Experiment 3.

Adelson, E. H. (2000). Lightness Perception and Lightness Illusions. In M. Gazzaniga (Ed.), *The New Cognitive Neurosciences, 2nd edition* (pp. 339-351). Cambridge, MA: MIT Press.

American Society for Testing and Materials. (2017 ). Standard test method for luminous reflectance factor of acoustical materials by use of integrating-sphere reflectometers. *Renovations of Center for Historic Preservation, 98(A)*, E1477.

Arend, L. E., & Goldstein, R. (1987). Simultaneous constancy, lightness, and brightness. *J Opt Soc Am A, 4*(12), 2281-2285. doi:10.1364/josaa.4.002281

Arend, L. E., & Spehar, B. (1993a). Lightness, brightness, and brightness contrast: 1. Illuminance variation. *Percept Psychophys, 54*(4), 446-456. doi:10.3758/bf03211767

Arend, L. E., & Spehar, B. (1993b). Lightness, brightness, and brightness contrast: 2. Reflectance variation. *Percept Psychophys, 54*(4), 457-468. doi:10.3758/bf03211768

Brainard, D. H. (1989). Calibration of a computer controlled color monitor. *Color Research & Application, 14*(1), 23-34.

Brainard, D. H. (2015). Color and the Cone Mosaic. *Annu Rev Vis Sci, 1*, 519-546. doi:10.1146/annurev-vision-082114-035341

Brown, R. O., & MacLeod, D. I. (1997). Color appearance depends on the variance of surround colors. *Curr Biol, 7*(11), 844-849. doi:10.1016/s0960-9822(06)00372-1

Fechner, G. T. (1966). *Elements of psychophysics* (Vol. 1). New York: Holt, Rinehart and Winston.

Gilchrist, A. L. (1977). Perceived lightness depends on perceived spatial arrangement. *Science, 195*(4274), 185-187. doi:10.1126/science.831266

Green, D. M., & Swets, J. A. (1996). *Signal detection theory and psychophysics* (Vol. 1). New York: Wiley.

Heasly, B. S., Cottaris, N. P., Lichtman, D. P., Xiao, B., & Brainard, D. H. (2014). RenderToolbox3: MATLAB tools that facilitate physically based stimulus rendering for vision research. *J Vis, 14*(2). doi:10.1167/14.2.6

Hecht, S., Shlaer, S., & Pirenne, M. H. (1942). Energy, Quanta, and Vision. *J Gen Physiol, 25*(6), 819-840. doi:10.1085/jgp.25.6.819

Hillis, J. M., & Brainard, D. H. (2005). Do common mechanisms of adaptation mediate color discrimination and appearance? Uniform backgrounds. *J Opt Soc Am A Opt Image Sci Vis, 22*(10), 2090-2106. doi:10.1364/josaa.22.002090

Hillis, J. M., & Brainard, D. H. (2007). Do common mechanisms of adaptation mediate color discrimination and appearance? Contrast adaptation. *J Opt Soc Am A Opt Image Sci Vis, 24*(8), 2122-2133. doi:10.1364/josaa.24.002122

Jakob, W. (2010). Mitsuba renderer. doi:<http://www.mitsuba-renderer.org>

Kelly, K. L., Gibson, K. S., & Nickerson, D. (1943). Tristimulus specification of the Munsell book of color from spectrophoto-metric measurements. *Journal of the Optical Society of America, 33*(7), 355-376.

Kingdom, F. A. (2011). Lightness, brightness and transparency: a quarter century of new ideas, captivating demonstrations and unrelenting controversy. *Vision Res, 51*(7), 652-673. doi:10.1016/j.visres.2010.09.012

Lotto, R. B., & Purves, D. (1999). The effects of color on brightness. *Nat Neurosci, 2*(11), 1010-1014. doi:10.1038/14808

Marimont, D. H., & Wandell, B. A. (1994). Matching color images: the effects of axial chromatic aberration. *Journal of the Optical Society of America A, 11*(12), 3113-3122.

Nachmias, J., & Sansbury, R. V. (1974). Letter: Grating contrast: discrimination may be better than detection. *Vision Res, 14*(10), 1039-1042. doi:10.1016/0042-6989(74)90175-8

Parker, A. J., & Newsome, W. T. (1998). Sense and the single neuron: probing the physiology of perception. *Annu Rev Neurosci, 21*, 227-277. doi:10.1146/annurev.neuro.21.1.227

Pelli, D. G., & Farell, B. (1999). Why use noise? *Journal of the Optical Society of America A, 16*(3), 647-653.

Prins, N., & Kingdom, F. A. A. (2018). Applying the Model-Comparison Approach to Test Specific Research Hypotheses in Psychophysical Research Using the Palamedes Toolbox. *Frontiers in Psychology, 9*, 1250. doi:doi: 10.3389/fpsyg.2018.01250

S, I. (1977). Tests for Colour-Blindness. *Tokyo: Kanehara Shuppen Company, Ltd.*

Singh, V., Cottaris, N. P., Heasly, B. S., Brainard, D. H., & Burge, J. (2018). Computational luminance constancy from naturalistic images. *J Vis, 18*(13), 19. doi:10.1167/18.13.19

Stockman, A., & Brainard, D. H. (2010). Color vision mechanisms. *OSA Handbook of Optics,* , 1-11.

Teller, D. Y. (1984). Linking propositions. *Vision Res, 24*(10), 1233-1246. doi:10.1016/0042-6989(84)90178-0

Toscani, M., Zdravkovic, S., & Gegenfurtner, K. R. (2016). Lightness perception for surfaces moving through different illumination levels. *J Vis, 16*(15), 21. doi:10.1167/16.15.21

Vrhel, M. J., Gershon, R., & Iwan, L. S. (1994). Measurement and analysis of object reflectance spectra. *Color Research & Application, 19*(1), 4-9.