Title: Equivalent noise characterization of human lightness constancy

Authors: Vijay Singh[1,2], Johannes Burge[2,3,4,5], David H. Brainard[2,3]

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 representational stability in the context of lightness, we introduce a threshold-based psychophysical paradigm. We measure how thresholds for discriminating the lightness of a target object (task-relevant property) in graphically-rendered naturalistic scenes are impacted by variation in the reflectance functions of background objects (task-irrelevant property). Our approach has roots in the equivalent noise paradigm. This paradigm relates signals to internal and external sources of noise, and it has been traditionally used to investigate contrast coding. We observe that, for low variation in background reflectance, the discrimination thresholds were nearly constant, indicating that observers' internal noise determines threshold in this regime. As the variation in background object reflectance increases, its effects start to dominate performance. We measure 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 linear receptive field model, based on Signal Detection Theory, which employs a single center-surround receptive field tailored to our stimulus set, 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 provides stable perceptual representations of the distal properties of objects. The computations that give rise to these representations start with the information in the proximal stimuli reaching the retinas. These proximal stimuli depend on the vagaries of the observer's particular viewpoint on the scene, on object-extrinsic properties of the scene (e.g. illumination), and on the intrinsic properties of the objects in the scene. The challenge for the visual system is to recover stable correlates of object-intrinsic properties across variation in other scene variables. Understanding the degree to which the visual system solves this challenge, and how it does so, is an important goal of vision science (Brascamp & Shevell, 2021; Helmholtz, 1896; Knill & Richards, 1996).

Here we consider the perceptual task of representing the reflectance of an 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 an object is the object's color appearance or, in the special case of achromatic objects, its lightness. Computing a stable color and lightness representation poses a challenge to the visual system because the retinal image of the object varies with the object's reflectance, the spectral irradiance of the illumination, and the position and pose of the object in the scene. The degree that the visual system succeeds at stabilizing its object color and lightness representation in the face of this variation determines the degree to which we say that the visual system achieves color and lightness constancy.

Color and lightness constancy have been studied extensively using psychophysical methods, in which observers report the color and lightness they perceive, across changes in the scene extrinsic to the judged object's reflectance. This literature tells us that under some viewing conditions, the visual system can achieve high degrees of constancy in the face of changes in illumination (Foster, 2011). This constancy is mediated by several different cues, including the mean and variance of the light reflected from the overall scene, from the immediate background of the object being judged, and from the more luminous regions of the scene (Brainard & Radonjić, 2014; Hurlbert, 2019; Smithson, 2005). A variety of theoretical frameworks have been developed, which provide various means for understanding how different cues are combined and how they shape the ultimate perceptual representations of object reflectance (see reviews cited earlier in this paragraph as well as Adelson, 2000; Brainard & Maloney, 2011; Gilchrist, 2006; Kingdom, 2011; Murray, 2021).

Objective psychophysical methods complement subjective measurements of appearance. These methods, which often involve determining threshold for discriminating changes along a specified dimension of stimulus variation, do not provide reports of what the stimulus looks like, but instead more directly assess the precision of the perceptual representation. Psychophysical threshold measurements are accompanied by a mature theory that can be used to link the measurements to properties of physiologically measured neural responses (Brindley, 1960; Green, 1996; Parker & Newsome, 1998; Teller, 1984). In addition, threshold measurements can be readily adapted for use with non-human subjects, since trial-by-trial reward can be provided based on whether each response is correct. It is less clear, however, how to apply threshold measurements to questions of perceptual constancy. One approach is to connect thresholds to appearance measurements, an approach which has its origins in Fechner's pioneering interpretation of Weber's Law (Fechner, 1966). The fundamental idea is that both thresholds and appearance are mediated by a common stimulus-response function whose properties depend on and can change with viewing context. Thresholds for making discriminations are related to the slope of the response function, with higher slopes leading to larger response changes for a fixed stimulus change and thus lower thresholds. Appearance, on the other hand, is related to the value of the response function, which provides the magnitude of the response. This approach holds promise (Hillis & Brainard, 2005, 2007b; Nachmias & Sansbury, 1974), but there are documented cases where the threshold measurements fail to account for appearance effects related to lightness constancy (Hillis & Brainard, 2007a). Another threshold-based approach to constancy is to study ability to detect a change in the confounding scene property (e.g., the

Commented [VS1]: Can we use English please?

Commented [BDH2]: Emailed Karl to see if his constancy work with Bloj is published.

illumination; Alvaro, Linhares, Moreira, Lillo, & Nascimento, 2017; Aston, Radonjić, Brainard, & Hurlbert, 2019; Pearce, Crichton, Mackiewicz, Finlayson, & Hurlbert, 2014; Radonjić et al., 2018; Radonjić et al., 2016), rather than a change in the object property of interest (e.g. surface reflectance). The goal here to measure the range of illumination changes over which the visual system's representation of object surface reflectance remains constant. How the results of measurements of this sort, which probe threshold-level illumination changes, relate to the stability of object appearance across larger illumination changes that occur in natural viewing has not been worked out (but see Weiss, Witzel, & Gegenfurtner, 2017).

Here we introduce a new approach to using a psychophysical threshold paradigm to draw inferences about perceptual constancy and its underlying mechanisms, based on measuring how thresholds for a task-relevant scene property are affected by variation in a task-irrelevant scene property. This approach is conceptually similar to studying how thresholds are affected by addition of spatially white or pink noise (Legge, Kersten, & Burgess, 1987; Pelli, 1990; Pelli & Farell, 1999), but with the noise introduced as variation in distal scene properties. In this paper, we apply this approach towards understanding lightness constancy in naturalistic graphically-rendered scenes, but the ideas are general. First, we measure human ability to discriminate the lightness of two objects in the absence of any object-extrinsic variation. Next, we measure how these discrimination thresholds change with the introduction of object-extrinsic variation in the form of variability in the colors (i.e. reflectance spectra) of background objects in the scene (Brown & MacLeod, 1997; Lotto & Purves, 1999). The discrimination threshold at each level of background variation quantifies the difficulty of the lightness discrimination task in each condition. 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 are 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 model rooted in Signal Detection Theory, that makes use of a single center-surround receptive field captures the essential features of the psychophysical data.

Commented [VS3]: Johannes did not understand the previous version of this sentence. This is my attempt to explain the idea.

Commented [BDH4]: Emailed Karl to see if he's published the illum discrim work that this paper refers to as under review.

Deleted: 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

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.

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 the 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 increase 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.

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 (T) with the variance in the internal noise (σ_i^2), the external noise (σ_e^2), and the covariance scalar (σ^2) as:

$$T^2 = T_0^2\ (\sigma_i^2 + \sigma^2\sigma_e^2) \quad (1)$$

where T_0 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 that is double that of threshold with no external variation indicates when the level of the external noise matches the level of internal noise.

To relate the SDT model directly to the stimuli used in our experiments, we developed a version of the SDT model based on a single-channel linear receptive field (see Methods: Linear Receptive Field Model; LINRF). This version of the model calculates the response of a linear receptive field to an image to instantiate the internal representation for the LRV postulated in the SDT model. This representation is then used in a simulation of the 2AFC paradigm to estimate model threshold. The receptive field model has the advantage that it can incorporate the Poisson noise that perturbs cone photoreceptor isomerizations as well as account for the truncation of surface reflectances to the range 0 to 1 in our model of natural surface reflectances.

Figure 4 shows the fit of the SDT model and the LINRF 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 LINRF model provides a better fit, because this model accounts for 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).

The model fits provide estimates of internal and external noise for our task, in the stim of the human observers in this task. Figure 6 provides the estimates of the internal and external noise standard deviations (quantities σ_i and σ_{e0}, see Methods) for the SDT model and the LINRF model.

The estimates of the internal noise are similar for the two models, which makes sense since for zero external noise the model formulations converge; the Poisson noise included in the LINRF model does not typically limit human discrimination performance at daylight light levels (Banks, Geisler, & Bennett, 1987; Cottaris, Jiang, Ding, Wandell, & Brainard, 2019). The mean values of the internal noise standard deviation are close to the values obtained by fitting the mean data (SDT model: mean value of internal noise standard deviation 0.0256, value from fit to mean data 0.0256; LINRF model: mean value of internal noise standard deviation 0.0250, value from fit to mean data, 0.0250).

The estimates of external noise are higher for the LINRF model than for the SDT model (SDT model: mean value of external noise standard deviation 0.0290, value from fit to mean data 0.0294; LINRF model: mean value of external noise standard deviation 0.0421, value from fit to mean data, 0.0429). This is consistent with the observation that the SDT model underestimates the rise in thresholds with

increasing covariance scalar, while this rise is captured accurately by the LINRF model, presumably because the latter takes into account the truncation implemented in the Gaussian model of natural surface reflectance variation. If we focus on the estimates from the LINRF model fit to the mean data, σ_{e0} is larger by a factor of ~1.7 than σ_i. To the extent that our model of natural surface reflectance is accurate, this tells us that as measured by our paradigm, the visual systems stabilization of its lightness variation against variation in background surface reflectance is within a factor of two of the limits imposed by the intrinsic precision of that representation.

DISCUSSION

The perceived lightness of an object can depend on the scene in which it lies. Stabilization of the lightness representation against variation in scene properties extrinsic to the object's surface reflectance is referred to as lightness constancy. In this paper, we introduced an objective psychophysical approach to characterizing lightness constancy, based on measuring how lightness discrimination thresholds vary with experimentally introduced variation in scene properties extrinsic to the object's reflectance. Specifically, we studied how lightness discrimination thresholds are impacted by variation in the reflectance of the background objects in graphically rendered naturalistic scenes. Our results (Figures 4 and 5) 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 observer. As the amount of background color variation increases, the effect of external variation in the stimuli on observers' representation of object lightness starts dominating that of the internal noise, and discrimination thresholds increase. We analyzed the data using a modeling approach similar to that used previously in the study of how externally added noise affects contrast detection (Legge et al., 1987; Pelli, 1990; Pelli & Farell, 1999). This approach allows us to relate the effect of background surface variation to the intrinsic precision of the lightness representation. We find that the effect of the external variability introduced by variation of background surface reflectances in naturalistic images is within a factor of two of the intrinsic precision of the lightness representation. More generally, our work provides a method to quantify the effect of variations in task-irrelevant properties on the perception of task-relevant property. Below we discuss some aspects of our approach and findings.

Spatial and chromatic properties of the stimuli. We used small image patches in our study, an important difference between our stimuli and natural viewing. In this initial deployment of our paradigm, we thus focus on effects of the background that are relatively nearby the test object, and which are likely mediated by relatively small populations of neurons. The use of small image patches is not a necessary requirement of our paradigm, and extending the work to larger patches is natural direction. In addition to using small patches, we did not vary the spatial structure of the array of objects in the rendered scene. Manipulating spatial structure may provide a way to use our paradigm to measure the spatial tuning of the mechanism(s) mediating the background effect. This approach is loosely analogous to how manipulating the structure of contrast noise may be used to examine the tuning of mechanisms supporting the detection of contrast-defined targets (Henning, Hertz, & Hinton, 1981; Losada & Mullen, 1995; Nachmias, 1999; Rovamo, Franssila, & Nasanen, 1992; Rovamo, Raninen, & Donner, 1999). Similarly, it may be possible to manipulate the chromatic structure of the variation in background surface reflectances with the goal of understanding the chromatic tuning of the background effect. This would again be analogous to how noise-based approaches have been used to characterize chromatic tuning of mechanisms that support the detection of chromatically-defined targets (Gegenfurtner & Kiper, 1992; Giulianini & Eskew, 1998; Monaci, Menegaz, Süsstrunk, & Knoblauch, 2004; Sankeralli & Mullen, 1997).

Commented [BDH6]: These might not be the perfect set of references, but they'll get us started.

Link between thresholds and appearance. The technique developed here probes the constancy of a perceptual representation of a task-relevant variable (here perceived object lightness) by measuring how variation in a task-irrelevant scene variable (here the reflectance of the background surfaces) elevate thresholds for detecting changes in the task-relevant variable. As with other threshold-based methods for approaching the stability of object appearance (see Introduction), it is not known the extent to which the

results of our technique may be used to predict the task-relevant stability of object appearance, as judged using subjective methods, across changes in a task-irrelevant scene variable. Experiments that explore this link are of considerable interest.

Applications to understanding neural mechanisms. A longstanding goal of vision science is to connect psychophysical performance to its underlying neural mechanisms. For probing mechanisms that mediate perceptual constancies, our paradigm has the attractive property that there is a well-defined correct answer on each trial, so that performance-contingent rewards can be provided to animal subjects. There are analytical methods for linking psychophysical discrimination performance to the response properties of neural populations that can be applied to studies like those performed here (Cohen & Maunsell, 2011; Ruff & Cohen, 2019; Salzman & Newsome, 1994; Shadlen, Britten, Newsome, & Movshon, 1996). Complementing these measurements with normative analyses that leverage receptive fields that encode the optimal stimulus features for specific tasks and that optimally decode the responses of these receptive fields into estimates (or forced-choice responses) will help enrich our understanding of the links between sensory-perceptual processing, neural computation, and psychophysical performance (Burge, 2020; Burge & Jaini, 2017; Geisler, Najemnik, & Ing, 2009; Jaini & Burge, 2017). There have already been some successes of this approach in the domains of blur, binocular disparity, and speed estimation in naturalistic images (Burge & Geisler, 2011, 2014, 2015; Chin & Burge, 2020). We are excited about the potential of our paradigm to be adopted to provide rigorous quantitative insights about the sensory-perceptual processing and the neural computations underlying color constancy in particular, and perceptual constancy more generally.

Model of natural surface reflectances. We used a statistical model of naturally-occurring surface reflectances to determine the distribution from which we sampled the background surface reflectance functions. This model was developed in our earlier work (see also Brainard & Freeman, 1997; Singh et al., 2018; Zhang & Brainard, 2004). This model is based on measurements of surface reflectance functions of the Munsell papers (Kelly et al., 1943) as well as surfaces characterized by Vrhel (1994). Although we view this as a reasonable model, it is important to note that the quantitative relation we measured between the magnitude of internal noise and the effect of external noise for natural surface reflectances depends on our choice of surface reflectance model. If the model overestimates the variation in natural surfaces, the effect of external noise for such variation is less than we estimated. Conversely, if the model underestimates the variation in natural surfaces. We discuss elsewhere other approaches to modeling naturally occurring surface variations, and present limitations on those methods (Singh et al., 2018). Future refinement of surface reflectance models could be used in conjunction with the linear receptive field model we report here, to refine the estimate of the effect of naturally occurring background variation on object lightness perception.

Rule of combination. In the present work, we considered variation in only a single task-irrelevant variable. In natural scenes, there are many task-irrelevant variables. In the case of judging object lightness, these include object-extrinsic factors such as the scene illumination, the position and 3D orientation of the target object in the scene, the viewpoint from which the object is viewed, and various object-intrinsic factors like its shape and size. Variation in each of the factors could in principle elevate thresholds for discriminating object lightness. Our paradigm allows characterization of the effect of these task-irrelevant variables and quantifies that effect for each such variable in the same internal-noise referred units. One potentially important future direction is to measure the combined effect of simultaneous variation of multiple task-irrelevant variables, and to test hypotheses about the rules of combination.

Commented [BDH7]: Work on best refs for here.

Commented [JDB8]: A contrast hadn't really been drawn between task-relevant and task-irrelevant variation… so I scrapped it. Feel free to add back in. But we'd need to flesh out a bit more. Didn't read well as it was.

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 method for extracting threshold from the data were preregistered before the 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 deviation from the pre-registered plan for pre-registered 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 would be excluded if their mean threshold for the last two acquisitions run in the practice session exceeded 0.025. After collecting data from 8 naive observers, we concluded that this criterion was too strict as only one observer met the criterion. Hence, we increased exclusion threshold from 0.025 to 0.030. The pre-registered plans also indicated that each image would be presented for 500ms, but in the event we shortened this to 250ms.

We followed the procedure described in the pre-registration document to extract threshold from the data. The document also indicated that the primary data feature of interest was the dependence of threshold on the covariance scalar and predicted that thresholds would increase with increasing background variability. The quantitative models of the data, however, were 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

4.66cm x 4.66cm (3.56° x 3.56°). 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 gamma functions for each primary were determined from measurements of the spectral power distribution for each primary at 26 equally spaced input values for that primary, in the range [0, 1] where 1 corresponds to the maximum value of the allowed input and 0 corresponds to no input. These gamma functions as well as the light emitted by the monitor for an input of 0 were accounted for in the stimulus display procedures (Brainard, Pelli, & Robson, 2002). The spectral 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 used check the linearity of the display. The maximum absolute deviation of the x-y chromaticity between the measured values and those predicted based on linearity was 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 (Ishihara, 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 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 Details for 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 T^2, -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) (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 (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 (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 multivariate normal distribution. Reflectance spectra for the objects in the scene were generated using random sampling from this multivariate 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 was multiplied by a fixed scalar with an arbitrarily chosen value of 5 to get the illuminant spectrum. This spectrum was used for all light sources in the visual scene. We then scaled the rendered image dataset so that the maximum linear monitor input value (across the entire dataset) needed to render the images on the display was 0.9.

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 signal detection theory (Green, 1996). We model the visual response to the target object in each image by a univariate internal representation denoted by the variable z. This variable depends on the image and is perturbed by noise. We assume that for any fixed image, z is a Gaussian distributed random variable whose mean depends on the target object LRF. For each image, we assume that z 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 z for a fixed image as the internal noise and denote its variance as σ_i^2. For

each trial of the experiment, z takes on two values, z_s and z_c, 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 z for each pair of LRFs, μ_s and μ_c, and on the value of σ_i^2. 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 z 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 σ_e^2. We assume that σ_e^2 depends on the experimentally chosen covariance scalar, but not on the target sphere LRF. Thus, the distributions of z_s and z_c, for a particular choice of target standard and comparison LRF and covariance scalar, are given by $P(z_s) = N(\mu_s, \sigma_t)$ and $P(z_c) = N(\mu_c, \sigma_t)$. Here μ_s is the mean value of the internal representation to the standard image and μ_c is the mean value of the internal representation to the comparison image. The total variance σ_t is given as $\sigma_t^2 = \sigma_i^2 + \sigma_e^2$, where σ_i^2 and σ_e^2 are the variance of the internal and external noise.

In the standard formulation of Signal Detection Theory for a 2AFC task, the observer makes their decision based on a comparison of z_s and z_c, choosing the interval with the higher value of z. The observer's sensitivity depends on the mean values and the variance and is captured by the quantity d-prime given as $d' = (\mu_c - \mu_s)/\sigma_t$. This quantity ($d'$) measures the distance between the two distributions in standard deviations units. Thus $d' = 0$ corresponds to an inability to distinguish between the standard and the comparison image. Larger values of d' indicate increasing discriminability.

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

$$(\mu_c - \mu_s) = d'\sigma_t = d'\sqrt{(\sigma_i^2 + \sigma_e^2)} \tag{2}$$

We further assume that the difference in mean value of the internal variable $(\mu_c - \mu_s)$ is proportional to the difference in the LRFs of the target object in the standard and comparison images (Δ_{LRF}). That is, $(\mu_c - \mu_s) = C\,\Delta_{\mathrm{LRF}}$, where C is the proportionality constant. Then we have,

$$\Delta_{\mathrm{LRF}} = \frac{d'}{C}\sqrt{(\sigma_i^2 + \sigma_e^2)} \tag{3}$$

When we measure threshold in a 2AFC task, we choose a criterion proportional correct and find the Δ_{LRF} that corresponds to that proportion correct. Our choice of 0.76 corresponds to $d' = 1$. In addition we can choose $C = 1$, in essence setting the units for z to match 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 multivariate normal distribution to generate the surface reflectance functions of the background objects.[1] To change the amount of external noise, we scaled the variance of the multinormal distribution by multiplying its covariance matrix with a scalar. Thus, for our experiments we can write:

[1] Here we neglect the effect of the fact that we truncated the distribution to enforce a requirement that reflectance at each wavelength lies between 0 and 1. We return to account for this below.

$$\Delta_{\mathrm{LRF}} = \sqrt{\sigma_i^2 + \sigma^2 \times \sigma_{e0}^2} \tag{4}$$

where σ^2 is the covariance scalar and σ_{e0}^2 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

$$\log\left(\Delta_{\mathrm{LRF}}^2\right) = \log\left(\sigma_i^2 + \sigma^2 \times \sigma_{e0}^2\right) \tag{5}$$

The equation above predicts that the form of threshold $\log\left(\Delta_{\mathrm{LRF}}^2\right)$ as a function of covariance scalar σ^2 should increase monotonically. For small values of σ^2 ($\sigma^2 \ll \sigma_i^2/\sigma_{e0}^2$), the threshold will approach a constant giving $\log\left(\Delta_{\mathrm{LRF}}^2\right) \sim \log\left(\sigma_i^2\right)$. For large values of σ^2 ($\sigma^2 \gg \sigma_i^2/\sigma_{e0}^2$), the quantity $\log\left(\Delta_{\mathrm{LRF}}^2\right)$ will approach a straight line with slope 1 in the $\log\left(\Delta_{\mathrm{LRF}}^2\right)$ versus $\log(\sigma^2)$ plot. Fitting the measurements with Equation 5 allows us to check whether the model describes the data, as well as, to determine the two parameters σ_i^2 and σ_{e0}^2. 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 σ_{e0}^2 quantifies how much the variation in background surface reflectance intrudes on the internal representation z that mediates the lightness discrimination task, in a manner that may be compared to the intrinsic precision of that representation specified by σ_i^2.

Linear Receptive Field Model

When the external noise added to the images is characterized by a multivariate Gaussian, a simple linear receptive field (LINRF) 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 I, and the receptive field by the column vector R. The entries of I are the radiant power emitted by the monitor at each image location. The entries of R are the corresponding sensitivities of the linear receptive field to each entry of I. The response of the receptive field is given as $r_i = R^T I + \eta_i$, where η_i is a random variable representing a draw of zero mean internal noise (variance σ_{ri}^2) in the receptive field response for a fixed image. We assume that σ_{ri}^2 is independent of I.

Denote I_{s0} and I_{c0} as the standard and comparison images without external noise. External Gaussian noise is added to both I_{s0} and I_{c0}, with covariance matrix Σ_e. 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

$$r_{ic} = R^T\left(I_{c0} + \eta_e\right) + \eta_i = R^T I_{c0} + \eta \tag{6}$$

$$r_{is} = R^T\left(I_{s0} + \eta_e\right) + \eta_i = R^T I_{s0} + \eta. \tag{7}$$

Here η_e is a random variable representing a draw of external noise, η_i represents the internal noise, and η 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, η is Gaussian with variance

$$\sigma_\eta^2 = \left(\sigma_{ri}^2 + R^T \Sigma_e R\right). \quad (8)$$

The mean difference between the receptive field response to the comparison and the standard image is given by $(\mu_c - \mu_s) = R^T (I_{c0} - I_{s0}) = C' \Delta_{\mathrm{LRF}}$. Here I_{s0} and I_{c0} are the standard and comparison images without external noise added, C' is a constant, and Δ_{LRF} is as defined is the SDT section above.

We associate the linear receptive field response with the internal representation z 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 of the SDT model, we have

$$\Delta_{\mathrm{LRF}} = \frac{d'}{C'} \sqrt{\sigma_{ri}^2 + \sigma^2 \times (R^T \Sigma_{e0} R)} \quad (9)$$

where we have introduced the covariance scalar σ^2 in the term corresponding to the variance of the external noise and where Σ_{e0} denotes the covariance matrix of the external noise corresponding to the level of variation in natural images. Comparing to relation derived in the SDT model (Equation 4), we see that this is the same functional form for the relation between Δ_{LRF} and σ^2 as derived there, where we associate $\sigma_i^2 = \frac{\sigma_{ri}^2}{(C')^2}$ and $\sigma_{e0}^2 = \frac{(R^T \Sigma_{e0} R)}{(C')^2}$.

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 σ_i^2. This procedure is described in more detail below. Once the fitting procedure (described below) establishes R and σ_i^2 that best account for the data, we then find σ_{e0}^2 directly by passing the images corresponding to $\sigma^2 = 1$ through the receptive field and finding the resulting variance.

SDT Model Fit

The theory of signal detection model was fit to the threshold versus covariance scalar data to obtain the parameters σ_i^2 and σ_e^2. The parameters were obtained by minimizing the mean squared error between the measured and predicted threshold using the MATLAB function *fmincon*.

Linear Receptive Field Model Fit

We fit the LINRF model using a simulation approach. We used simulation for two reasons. First, it allowed us to incorporate a model of the early visual system into the computations. Second, it provides a way to account for truncation in the Gaussian model of natural surface reflectances. The truncation occurs because we require that surface reflectance at each wavelength lie between 0 and 1.

The model of early visual system was as described by Singh et al. (2018). The model was implemented using the software infrastructure provided by ISETBio (ISETBio; isetbio.org; Cottaris et al., 2019). It incorporated typical optical blurring, axial chromatic aberration (Marimont & Wandell, 1994), and spatial sampling by the mosaic of 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 (CIE, 2007) as implemented in ISETBio was used to

obtain LMS cone fundamentals. Cone excitations were calculated as the number of photopigment isomerizations in a 100ms integration time, and included simulation of the Poisson variability of the isomerizations (Rodieck, 1998). The cone isomerizations were demosaiced using linear interpolation to estimate LMS isomerization images. Further, the isomerizations of each cone class was normalized by the summed (over wavelength) quantal efficiency of the corresponding cone class, to make the magnitude of the signals from the three cone classes similar to each other. This normalization occurred after incorporation of Poisson noise and did not affect the signal-to-noise ratio of the signals from the different cone classes.

The dot product of the LMS isomerization images was taken with a simple center-surround linear receptive field. The receptive field was square in shape to match the image size. Its center was a circle of radius equal to the size and at the location of the target object in the image. 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 $\mathrm{v_c} = 1$, and each region of the surround had sensitivity denoted by $\mathrm{v_s}$. The RF was the same for each of the three cone classes. The RF response was taken as the sum of the L, M and S RF component responses. Gaussian internal noise with zero mean was added to the resulting dot product. The variance of the internal noise (σ_{ri}) and the value of the RF surround ($\mathrm{v_s}$) were the two parameters of the model.

The threshold predictions of the LINRF model for any choice of model parameters were obtained using simulation of a two-interval force choice paradigm similar to the experiment. For each trial, we randomly sampled a standard image and a comparison image from our dataset. We obtained the response of the receptive field (noise-added dot product) to the images and compared them to determine the simulated choice on that trial. This process was repeated 10,000 times for each of the 11 comparison LRF levels. The proportion comparison chosen data was used to get the psychometric function and the threshold of discrimination, similar to the method used for human 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 data being fit and the computational model for a large set of values of the two model parameters: the variance of the decision noise (σ_{ri}) and the value of the RF surround ($\mathrm{v_s}$). 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 estimate the parameters with lowest mean square error. These parameters were then used to estimate the internal and external noise standard deviation of the LINRF model using the relations: $\sigma_i^2 = \frac{\sigma_{ri}^2}{(C')^2}$ and $\sigma_{e0}^2 = \frac{(R^T \Sigma_{e0} R)}{(C')^2}$ as explained above, where the constant C' was obtained using the relationship $R^T(I_{c0} - I_{s0}) = C'\Delta_{\mathrm{LRF}}$.

The best parameters and the internal and external noise standard deviation were estimated separately for the mean observer and the individual observers.

Code and Data Availability

Observers' response in the psychophysics task and their thresholds are provided as supplementary information (SI). The SI also provides the MATLAB scripts to generate Figures 2, 4, 5 and 6 and the scripts to get thresholds of the LINRF model. The retinal images are provided as .mat files in a zip folder. The SI is available at: https://github.com/vijaysoophie/EquivalentNoisePaper

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. R_{N-1} indicates the time of the observer's response for the $(N-1)^{th}$ trial. The N^{th} trial begins 250ms after that response (Inter Trial Interval, ITI). The N^{th} 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 R_N 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) $T^2 = T_0^2\left(\sigma_i^2 + \sigma^2\sigma_e^2\right)$ with $T_0 = 1$ (red curve). The best fit parameters are indicated in the legend. The threshold of the linear receptive field (LINRF) model was estimated at 10 logarithmically spaced values of the covariance scalar (black squares). The black smooth curve is a smooth fit to these points of the

functional form $\log_{10} T^2 = a + b^{(x+c)^d}$ where $x = \log_{10} \sigma^2$ and a, b, c and d are parameters adjusted in the fit.

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 LINRF 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 (LINRF) model. While the internal noise estimates are consistent over the two models, the external noise estimated by the LINRF 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 to 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 ($\sigma^2 = 1$). The thresholds for this condition are also provided for comparison. Two observers from Experiment 2 also participated in Experiment 3.

REFERENCES

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.

Alvaro, L., Linhares, J. M. M., Moreira, H., Lillo, J., & Nascimento, S. M. C. (2017). Robust colour constancy in red-green dichromats. *PLoS ONE, 12*(6), e0180310.
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.
Aston, S., Radonjić, A., Brainard, D. H., & Hurlbert, A. C. (2019). Illumination discrimination for chromatically biased illuminations: implications for colour constancy. *Journal of Vision, 19*(30:15).
Banks, M. S., Geisler, W. S., & Bennett, P. J. (1987). The physical limits of grating visibility. *Vision Research, 27*(11), 1915-1924.
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. *Annual Review of Vision Science, 1*, 519-546.
Brainard, D. H., & Freeman, W. T. (1997). Bayesian color constancy. *Journal of the Optical Society of America A, 14*(7), 1393-1411.
Brainard, D. H., & Maloney, L. T. (2011). Surface color perception and equivalent illumination models. *Journal of Vision, 11*(5).
Brainard, D. H., Pelli, D. G., & Robson, T. (2002). Display characterization. In J. P. Hornak (Ed.), *Encylopedia of Imaging Science and Technology* (pp. 172-188). New York: Wiley.
Brainard, D. H., & Radonjić, A. (2014). Color constancy. *The New Visual Neurosciences, 1*, 545–556.
Brascamp, J. W., & Shevell, S. K. (2021). The certainty of ambiguity in visual neural representations. *Annual Review of Vision Science, in press*.
Brindley, G. S. (1960). *Physiology of the Retina and the Visual Pathway*. London: Arnold.
Brown, R. O., & MacLeod, D. I. A. (1997). Color appearance depends on the variance of surround colors. *Current Biology, 7*, 844-849.
Burge, J. (2020). Image-computable ideal observers for tasks with natural stimuli. *Annual Review of Neuroscience, 6*, 491-517.
Burge, J., & Geisler, W. S. (2011). Optimal defocus estimation in individual natural images. *Proceedings of the National Academy of Sciences, 108*(40), 16849-16854.
Burge, J., & Geisler, W. S. (2014). Optimal disparity estimation in natural stereo images. *Journal of Vision, 14*(2).
Burge, J., & Geisler, W. S. (2015). Optimal speed estimation in natural image movies predicts human performance. *Nature Communications, 6*, 7900.
Burge, J., & Jaini, P. (2017). Accuracy maximization analysis for sensory-perceptual tasks: computational improvements, filter robustness, and coding advantages for scaled additive noise. *PLoS Computational Biology, 13*(2), e1005281.
Chin, B. M., & Burge, J. (2020). Predicting the partition of behavioral variability in speed perception with naturalistic stimuli. *Journal of Neuroscience, 40*(4), 864-879.
CIE. (2007). *Fundamental chromaticity diagram with physiological axes – Parts 1 and 2. Technical Report 170-1*. Vienna: Central Bureau of the Commission Internationale de l' Éclairage.
Cohen, M. R., & Maunsell, J. H. (2011). Using neuronal populations to study the mechanisms underlying spatial and feature attention. *Neuron, 70*(6), 1192-1204.

Cottaris, N. P., Jiang, H., Ding, X., Wandell, B. A., & Brainard, D. H. (2019). A computational-observer model of spatial contrast sensitivity: Effects of wave-front-based optics, cone-mosaic structure, and inference engine. *Journal of Vision, 19*(4), 8.
Fechner, G. T. (1966). *Elements of Psychophysics*. New York: Holt, Rinehart and Winston.
Foster, D. H. (2011). Color constancy. *Vision Research, 51*(7), 674-700.
Gegenfurtner, K., & Kiper, D. C. (1992). Contrast detection in luminance and chromatic noise. *Journal of the Optical Society of America A, 9*(11), 1880-1888.
Geisler, W. S., Najemnik, J., & Ing, A. D. (2009). Optimal stimulus encoders for natural tasks. *Journal of Vision, 9*(13), 17 11-16.
Gilchrist, A. L. (2006). *Seeing Black and White*. Oxford: Oxford University Press.
Giulianini, F., & Eskew, R. T., Jr. (1998). Chromatic masking in the (DL/L, DM/M) plane of cone-contrast space reveals only two detection mechanisms. *Vision Research, 38*, 3913-3926.
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. *Journal of Vision, 14*(2).
Helmholtz, H. (1896). *Physiological Optics*. New York: Dover Publications, Inc.
Henning, G. B., Hertz, B. G., & Hinton, J. L. (1981). Effects of different hypothetical detection mechanisms on the shape of spatial-frequency filters inferred from masking experiments: I. Noise masks. *Journal of the Optical Society of America, 71*(5), 574-581.
Hillis, J. M., & Brainard, D. H. (2005). Do common mechanisms of adaptation mediate color discrimination and appearance? Uniform backgrounds. *Journal of the Optical Society of America A, 22*(10), 2090-2106.
Hillis, J. M., & Brainard, D. H. (2007a). Distinct mechanisms mediate visual detection and identification. *Current Biology, 17*(19), 1714-1719.
Hillis, J. M., & Brainard, D. H. (2007b). Do common mechanisms of adaptation mediate color discrimination and appearance? Contrast adaptation. *Journal of the Optical Society of America A, 24*(8), 2122-2133.
Hurlbert, A. (2019). Challenges to color constancy in a contemporary light. *Current Opinion in Behavioral Sciences, 30*, 186-193.
Ishihara, S. (1977). Tests for colour-blindness. *Tokyo: Kanehara Shuppen Company, Ltd.*
Jaini, P., & Burge, J. (2017). Linking normative models of natural tasks to descriptive models of neural response. *Journal of Vision, 17*(12), 16.
Jakob, W. (2010). Mitsuba Renderer.
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 Research, 51*(7), 652-673.
Knill, D. C., & Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge: Cambridge University Press.
Legge, G. E., Kersten, D., & Burgess, A. E. (1987). Contrast discrimination in noise. *Journal of the Optical Society of America A, 4*(2), 391-404.

Losada, M. A., & Mullen, K. T. (1995). Color and luminance spatial tuning estimated by noise masking in the absence of off-frequency looking. *Journal of the Optical Society of America A, 12*(2), 250-260.
Lotto, R. B., & Purves, D. (1999). The effects of color on brightness. *Nature Neuroscience, 2*(11), 1010-1014.
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.
Monaci, G., Menegaz, G., Süsstrunk, S., & Knoblauch, K. (2004). Chromatic contrast detection in spatial chromatic noise. *Visual Neuroscience, 21*, 291-294.
Murray, R. F. (2021). Lightness perception in complex scenes. *Annual Review of Vision Science, in press*.
Nachmias, J. (1999). How is a grating detected on a narrowband noise masker? *Vision Research, 39*(6), 1133-1142.
Nachmias, J., & Sansbury, R. V. (1974). Grating contrast: discrimination may be better than detection. *Vision Research, 14*(10), 1039–1042.
Parker, A. J., & Newsome, W. T. (1998). Sense and the single neuron: probing the physiology of perception. *Annual Review of Neuroscience, 21*(1), 227-277.
Pearce, B., Crichton, S., Mackiewicz, M., Finlayson, G. D., & Hurlbert, A. (2014). Chromatic illumination discrimination ability reveals that human colour constancy is optimised for blue daylight illuminations. *PLoS ONE 9*(2:e87989), e87989.
Pelli, D. G. (1990). The quantum efficiency of vision. In C. Blakemore (Ed.), *Vision: Coding and Efficiency* (pp. 3-24).
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 tesearch hypotheses in psychophysical research using the Palamedes toolbox. *Frontiers in Psychology, 9*, 1250.
Radonjić, A., Ding, X., Krieger, A., Aston, S., Hurlbert, A. C., & Brainard, D. H. (2018). Illumination discrimination in the absence of a fixed surface-reflectance layout. *Journal of Vision, 18*(5:11).
Radonjić, A., Pearce, B., Aston, S., Krieger, A., Dubin, H., Cottaris, N. P., et al. (2016). Illumination discrimination in real and simulated scenes. *Journal of Vision, 16*(11:2), 1-18.
Rodieck, R. W. (1998). *The First Steps in Seeing*. Sunderland, Mass.: Sinauer.
Rovamo, J., Franssila, R., & Nasanen, R. (1992). Contrast sensitivity as a function of spatial frequency, viewing distance and eccentricity with and without spatial noise. *Vision Research, 32*(4), 631-637.
Rovamo, J., Raninen, A., & Donner, K. (1999). The effects of temporal noise and retinal luminance on foveal flicker sensitivity. *Vision Research, 39*, 533-539.
Ruff, D. A., & Cohen, M. R. (2019). Simultaneous multi-area recordings suggest that attention improves performance by reshaping stimulus representations. *Nature Neuroscience, 22*(10), 1669-1676.
Salzman, C. D., & Newsome, W. T. (1994). Neural mechanisms for forming a perceptual decision. *Science, 264*, 231-237.

Sankeralli, M. J., & Mullen, K. T. (1997). Postreceptoral chromatic detection mechanisms revealed by noise masking in three-dimensional cone contrast space. *Journal of the Optical Society of America A, 14*(10), 2633-2646.
Shadlen, M. N., Britten, K. H., Newsome, W. T., & Movshon, J. A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience, 16*, 1486-1510.
Singh, V., Cottaris, N. P., Heasly, B. S., Brainard, D. H., & Burge, J. (2018). Computational luminance constancy from naturalistic images. *Journal of Vision, 18*(13), 19.
Smithson, H. E. (2005). Sensory, computational, and cognitive components of human color constancy. *Philosophical Transactions of the Royal Society of London. Series B, 360*(1458), 1329-1346.
Teller, D. Y. (1984). Linking propositions. *Vision Research, 24*(10), 1233-1246.
Vrhel, M. J., Gershon, R., & Iwan, L. S. (1994). Measurement and analysis of object reflectance spectra. *Color Research & Application, 19*(1), 4-9.
Weiss, D., Witzel, C., & Gegenfurtner, K. (2017). Determinants of colour constancy and the blue bias. *i-Perception, 8*(6), 204166951773963.
Zhang, X., & Brainard, D. H. (2004). *Bayesian color correction method for non-colorimetric digital image sensors.* Paper presented at the Color and Imaging Conference.