Title: Equivalent noise characterization of human lightness constancy

Authors: Vijay Singh, Johannes Burge, David H. Brainard

**Abstract:** A goal of visual perception is to provide stable representations of task-relevant scene properties (e.g. target object size, shape, 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). Four observers viewed computer-rendered images of a 1-degree sphere, within a 2-degree scene containing naturalistic background objects. The sphere’s reflectance was spectrally flat but varied in its overall level. On each trial, two images of the scene were presented in sequence and observers indicated which 250ms interval contained the sphere with higher reflectance. Across intervals, the reflectances of the background objects were randomized by sampling from a probabilistic model of naturally occurring surface reflectances. Discrimination thresholds were measured as a function of the amount of variability in the background object reflectance functions, which we controlled in a parametric fashion. This paradigm has roots in the use of contrast threshold versus noise (TvN) measurements to characterize the visual coding of contrast. 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, resulting in rising discrimination thresholds. The level of variability at which thresholds begin to rise quantifies the equivalent noise - the level of variation at which the task-irrelevant variable (background object reflectance) intrudes on the visual representation of the task-relevant variable (target object reflectance), to the same degree as the intrinsic variability of the internal representation. A computational model that uses a center-surround receptive field to estimate object lightness captures human behavior at this task. Our approach provides a novel method to characterize the effect of task-irrelevant scene variations on the perceptual representation of a task-relevant scene property.

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

**Introduction:** To support effective thought and action, vision needs to provide stable perceptual representations of the distal properties of objects, starting 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 thus 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 a central goal of vision science.

Here we consider the perceptual task of representing the reflectance of an achromatic object from the light reflected from the object to the eye. The corresponding perceptual representation is termed lightness – to put it another way, lightness is the perceptual correlate of the surface reflectance of achromatic objects. Providing a stable lightness representation challenges the visual system because the retinal irradiance of the image of the object varies both with the object’s overall reflectance but also with the irradiance of the illumination and the position and pose of the object in the scene. To the degree that the visual system successfully stabilizes the lightness representation against object-intrinsic variation, it is said to be lightness constant.

The perceptual representation of lightness has been extensively studied using appearance methods, in which observers report their subjective judgment of the lightness of objects.

It is related to the amount of light that is reflected by the object’s surface [1]. However, the light reflected from the object also depends on the context in which the object lies. It is well known that the perceived lightness of an object is affected by the background [2]. The manner in which variation in the background impacts the corresponding variation in the perceived lightness is not well understood.

Here, we empirically establish this relation for variation in the spectra of background objects. To do so, we measure the ability of human observers to discriminate the lightness of two objects. We study how discrimination thresholds change with variation in the color of objects in the background. We randomize the color of background objects by sampling from a statistical model based on the reflectance spectra of natural surfaces. We measure discrimination thresholds as a function of the amount of background spectral variation, which is controlled by a single model parameter. The effect of background spectral variation is quantified by the difficulty of the lightness discrimination task. We observe that as the variation in background color is increased, the 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 estimates the object lightness captures the essential features of human observers.

**Results:**

1. **Measurement of lightness discrimination thresholds**

We measured how variation in the reflectance spectra of background objects affect lightness discrimination thresholds. The psychophysical task is illustrated in Figure 1. We used a two-alternative forced-choice (TAFC) experiment to measure the thresholds. On each trial, observers viewed two images: a standard image and a comparison image (Figure 1a). The images were shown on a calibrated monitor for 250ms each, one after the other. A 250ms inter-stimulus interval (blank screen) was presented between the two images (Figure 1b; also Methods: Experimental Details). The images were computer graphics renderings of 3D scenes. Each image contained an achromatic spherical target object. The observer’s task was to report the image in which the target object was lighter. Across trials, we varied the luminous reflectance factor (LRF, [3]) of the sphere in the comparison image while keeping the LRF of the sphere in the standard image fixed. The LRF is the ratio of the luminance of a surface under a reference illuminant (here CIE D65; ref) to the luminance of the reference illuminant itself.

We recorded the proportion of times the observers chose the comparison image to have the lighter target object at 11 values of the target object LRF. Figure 2 shows a psychometric function from a typical human observer in our psychophysical task. We fit the proportion comparison chosen data with a cumulative Gaussian using maximum likelihood methods (See Methods: Psychometric Function). We defined the threshold 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 variation in reflectance of background objects**

To study the effect of variation in background on lightness discrimination thresholds, we varied the reflectance spectra of the background objects in the images by sampling them from a statistical model based on natural surface reflectance databases (See Methods: Reflectance and Illumination Spectra, [4, 5]). Briefly, a database of natural surface reflectance functions 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.

We measured threshold as a function of a scalar that multiplied the covariance matrix of the projection weight distribution. 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 from six logarithmically spaced values of the covariance scalar. Figure 3 shows particular image samples used in our psychophysical task. Discrimination thresholds were measured separately for each of the six values of the covariance scalar.

### 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 Signal Detection Theory (Figure 4; see Methods: Signal Detection Theory and Model Fit).

When the covariance scalar is 0, we conceptualize performance as limited by two factors (ref). One is internal variability in the observer’s representation of target object lightness. The other is the efficiency with which the observer’s decision processes make use of the information provided by this representation. Our experiments do not separate the relative contributions of these two conceptually distinct factors, and in the following we refer to them collectively as the observer’s internal noise for the lightness discrimination task.

As the covariance scalar increases, a third factor comes into play. This is the external variability introduced by the variation in background objects, to the extent that it impinges on the representation of target object lightness. At low values of the covariance scalar, the internal noise dominates the effect of the external variability and threshold remains roughly constant. At high vsalues of the covariance scalar, the effect of external variability limits performance, and thresholds rise with 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 observer shows the same basic pattern as the mean results, with constant thresholds across low values of the covariance scalar and then a rise of thresholds that is approximately linear on the log threshold squared versus log covariance plot. The most notable individual difference is in the slope of the rising limb measured functions.

1. **A computational model with center-surround receptive field captures human threshold increase**

We modeled human response in the psychophysical task using a computational observer model (See Methods: Model of visual-system; [6]). In this model, we first simulate the response of the early visual system to the images in our database. The model of the visual system incorporates optical blurring, axial chromatic aberration, spatial sampling of the cone mosaic and the Poisson nature of the photopigment isomerization. We simulate the response of the long (L), middle (M), and short (S) type photoreceptors of this model eye to each image in our image database. The rest of the visual system is modeled as a center surround receptive field. The RF consists of three channels, one each for the L, M, and S channels of the retinal image. The size of individual L, M and S channels of the RF was chosen to be equal to the extent of the retinal image. The center of the RF was a uniform circular patch the same size as the target object in the images. The surround was a uniform square patch of the size of the images except the target object. The mean response of the computational observer is estimated as the dot product of this center surround RF with the retinal images (simulated response of L, M, S photoreceptors to the images).

The response of the computational observer to the standard and comparison images were used in a two alternate forced choice paradigm, similar to the human psychophysics experiments, to obtain the threshold of the computational observer. To compare model thresholds with humans, we added a gaussian noise to the receptive response. The Gaussian noise had zero mean and its variance was proportional to the dot product of the receptive field with the image. The proportionality constant and the ratio of the RF surround to center were the two parameters of the model. These were chosen such that the mean square error between the threshold of the computational model and the mean human observer. We calculated the equivalent noise and the rate of threshold increase of the computational observer using the fit to Signal Detection Theory model (see Methods: Signal Detection Theory and Model Fit). These values compare well with the values of the mean human observer (Internal Noise Human = 0.0256, Model = 0.0257; External Noise Human = 0.0294, Model = 0.0305; Figure 4).

**Discussion:**

**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 designed three experiments. The main paper reports the findings of the third experiment (Experiment 3). The findings of the second experiment are reported as supplementary information. The first experiment (Experiment 1) was abandoned due to reasons explained below.

Experiment 1 aimed at the measurement of human object luminance discrimination thresholds using naturalistic computer graphics images. Luminance is the equivalent of lightness for chromatic objects [7]. In this experiment, the task was to discriminate two images based on the luminance of a target object. The hue and chroma of the target object were allowed to vary from image-to-image. This experiment was abandoned after performing preliminary threshold measurements on the experimenters. The reason was the large variability in luminance judgement due to chromatic aspects of the target object, even when the object-extrinsic variables were kept fixed. Since, our goal was to measure the effect of object extrinsic properties on lightness discrimination thresholds, to minimize the effect variations associated with the target object, we designed the next experiments with achromatic target objects.

Experiment 2 (see Supplementary Experiments) aimed at measurement of human object lightness discrimination thresholds under variation in the reflectance of object background. In this experiment, we measured lightness discrimination thresholds for three conditions: (1) Fixed background condition, where the background was fixed throughout all trials and intervals, (2) Between-trial background variation condition, where the reflectance of objects in the background was fixed for the two intervals of a trial, but varied randomly from trial-to-trial, and (3) Within-trial background variation condition, where the reflectance of objects in the background varied randomly in each trial and interval. This experiment established that lightness discrimination thresholds are higher for the condition when the two objects are being discriminated against different backgrounds on the same trial (Condition 3), as compared to the condition when the backgrounds are the same within trial (Condition 1 and 2). Trial-to-trial variability in background has little effect, if any. See Supplementary Experiments for details.

This paper describes Experiment 3 which was aimed at measurement of human object lightness discrimination thresholds as a function of the amount of variation in the reflectance of the object’s background.

The pre-registration document also specified the primary methods to analyze the data. It specified that the data will be analyzed separately for each subject by fitting a cumulative normal to the proportion comparison chosen data using the maximum likelihood method. The thresholds were to be extracted from the fit as the difference between object LRV at 0.7602 and 0.50 proportion comparison chosen. The subject thresholds at each level of background variability were to be measured three times and averaged. We predicted that thresholds would increase as the background variability increases.

**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 60 Hz, 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 gamepad (Logitech F310).

The observer’s head position was stabilized using chin cup and forehead rest (Headspot, UHCOTech, Houstion, 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). The spectral range of the spectroradiometer is 380-780nm at 4nm spacing, with an 8nm bandwidth and an accuracy of +/-2nm. To calibrate the monitor, we focused the spectroradiometer on a patch on the center of the monitor. The patch was of the size 4.8cm x 4.6cm (radiometer 75cm from screen, 3.67° by 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 at 26 values of the input 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 deviation of the x-y chromaticity and luminance between the applied and measured values was less than 1% for the calibration.

**Observer Recruitment and Exclusion:** The observers were recruited from the University of Pennsylvania and the local Philadelphia community and were compensated for their time. Before the start of measurements, observers were screened to have normal visual acuity (20/40 or better) and normal color vision (as tested with pseudo-isochromatic plates, [8]).

Observers were also screened for their 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 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 criteria 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 11 observers participated in the practice sessions (7 Female, 4 Male; age 18-56; mean age 30.4). Four of these met the criteria 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:** Our aim was to measure lightness discrimination thresholds for a target object as we varied the color of the objects in the background. We measured the 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 lightness. The lightness 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 choose this to be the standard daylight spectrum D65 (See Methods: Reflectance and Illumination Spectra). The geometry of the 3D scene and the spectral power distribution of the light sources were kept fixed.

**Experimental Details:** We used a two-interval forced choice procedure to measure the thresholds. We showed two images, one after the other, on a calibrated computer monitor and asked the observer to report the image 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 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 pregenerated image databases. The first five trials of each acquisition were moderate trials (as defined above) to acclimatize the observer to the experimental conditions. The responses for these five trials were not saved.

Each acquisition 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 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.

**Stimulus Presentation:** The images were presented on an LCD monitor. The monitor was located at a distance of 75cm from the observer. 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 0.25s (this was a deviation from the preregistration document, which specifies the presentation time as 0.5s), with an inter-stimulus interval of 0.25s and inter-trial interval of 0.25s. 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 0.25s (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; [9]). 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/), [10]). The images were rendered at 31 wavelengths equally spaced between 400 and 700nm. The images were rendered with the camera field of view of 17 degrees 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 [11] 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 ([7]). 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 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 also imposed a physical realizability condition on the generated samples by ensuring that the reflectance at each spectral frequency was within 0 and 1.

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 scale factor was applied to the power spectrum of all light sources in the visual scene.

**Experimental Procedure:** At the beginning of the first experimental session (the practice session), 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 instruments. 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.

**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 [12]. 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 0.7602 and 0.50 as obtained from the cumulative gaussian fit.

**Theory of Signal Detection:** Here we give a brief introduction to the theory of signal detection. For a more comprehensive discussion see [13]. We can model the response of the observers in the TAFC task as a decision based on an internal variable , that depends on the image presented on the screen. For simplicity, let us assume that for a given LRV of the target object in the image, the internal variable is Gaussian distributed random variable with a mean dependent on the target object LRV and the variance dependent on the total noise (including the internal noise, which is a combination of the noise in the visual representation of the image signal and the noise in the decision making process, and the external noise, which depends on the variability in the reflectance of background objects). In any trial of the experiment, the internal variable will take two values, and , whose distribution will be given as and for the standard and comparison images, respectively. Here is the mean response of the standard image and is the mean response of the comparison image. is the total variance, given as , where and are the variance of the internal and external noise, respectively, both assumed to be Gaussian distributed. In the standard formulation of Signal Detection Theory, the observer makes their decision based on the decision variable . The observer sets a criterion value to make the decision. If the decision variable is larger than the criterion , the observer chooses the comparison to have the larger LRV, otherwise they choose the standard image. 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 in terms of the standard deviation 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 :

Let’s assume that the difference in mean value of the internal variable is proportional to the difference in the LRVs of the target object in the standard and comparison images (), i.e., , where is the proportionality constant. Then,

given a set of measured values of for different values of external noise, one can use this relation to find the internal noise of the observer in units of the external noise.

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.

**Theory of Signal Detection for Model of visual system:** The computational model approximates the response of the visual system to the images as the output of a linear receptive field. The observer’s behavior is modeled as a noisy decision-making process based on this internal representation of the stimulus images by the visual system.

Let’s denote the stimulus image by the column vector , and the receptive field by the column vector . The response of the visual system is given as , where is a random variable representing a draw of the internal noise (mean zero, variance ) of the visual system. Neglecting noise, the difference in response between comparison and standard is , where and are the standard and comparison images without the noise. External noise is added to both and , with mean 0 and covariance matrix , where is the covariance matrix. After including noise, difference in the receptive field response is given as

Here is a random variable representing a draw of external noise in the image space, while represents the internal noise. This reduces to

where 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 mean 0 and variance

Assuming , following the arguments of the section Theory of Signal Detection, we have

where we have introduced the covariance scalar in the term corresponding to the variance in the external noise. Comparing to relation derived in the section Theory of Signal Detection, we have and .

**Model Fit:** Themodelwasfit to the threshold v/s 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.)

**Model of visual-system:** The response of the early visual system to the image database was estimated as described in [7]. The model incorporated typical optical blurring, axial chromatic aberration [14], and spatial sampling of the long (L), middle (M) and short (S) wavelength- sensitive cones [15]. 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 [16] was used to get the spectral sensitivities of the cones. The response of the cones was calculated as the number of photopigment isomerizations in 100 msec, including the Poisson nature of the isomerization [17]. 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 target object. The central region was taken to be positive and the surround was negative. Each point in the central region was at the same value (chosen to be unity) and the surround was at another value (). 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. A gaussian noise (representing noise in the decision-making process) with mean zero and variance proportional to the total mean RF response was added to the resulting dot product. The proportionality constant to the decision noise variance and the value of the RF surround () were the two parameters of the model.

To get the threshold of the computational model, we sampled random standard and comparison images from our dataset. The receptive field response (noise-added dot product) to the images were compared to predict the image with lighter target object. The proportion comparison chosen data was used to get the psychometric function and the threshold of discrimination. sixvalues of the 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 mean human observer and the computational model for a large set of values of the two model parameters: noise variance proportionality constant and RF surround. The resulting values were fit with a degree two polynomial of two variables. The fit function was used to get the parameters with lowest mean square error. These values were chosen as the parameters of the model (surround value -0.1293; noise proportionality constant 0.0904).

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

**Code and Data Availability:** Observers response in the psychophysics task and their thresholds are provided in the supplementary information. 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 (percent correct) was measured as a function of this difference to determine discrimination threshold. The reflectance functions of objects in the background could vary between standard and comparison (as illustrated) could also be varied trial to trial. 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 0.25s after that response (Inter Trial Interval, ITI). The Nth trial consists of two 0.25s stimulus presentation intervals with a 0.25s 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 time denoted by RN in the figure. The next trial begins 0.25s 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 CNSU\_0002 for scale factor 0.00 in the first experimental session for that observer. The point of subjective equality (PSE, LRF for 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 spectra.** The reflectance spectra of background objects were chosen from 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 variation in 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] LRF. 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 with . The predictions of the threshold of the computational model (black squares) and the corresponding model fit (black dashed line).

**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 smooth curves show the fit to the function , (). The parameters of the computational model were obtained separately for each observer by minimizing the mean square error between the mean observer thresholds and the predicted model thresholds for the six values of the covariance scalar.

**Supplementary Experiments:**

**Measurement of human object lightness discrimination thresholds under variation in object background:** This experiment was the precursor of the experiment performed in the main paper. It aimed to study 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. The methods were same as the experiment described in the main paper. We measured lightness discrimination threshold of four naïve human observers using a two-interval forced choice paradigm. The thresholds for were measured for three specific types of background variations (Figure S1). The reflectance spectra of the background objects were generated with 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.

In Condition 2 and 3, the light reflected from the target object varied from image to images (even at the same LRF level of the target object) because of secondary reflection of light coming from the background objects. 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.

**Observer Recruitment and Exclusion:** Same as for the main experiment, except the three acquisition of the practice session was performed with the images in Condition 1.

**Results:** Figure S2 shows the discrimination thresholds of four human observers for the five conditions studied in the supplementaryexperiment. 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 increase in threshold of the observers for Condition 3 and 3a as compared to Condition 1 (baseline) were 79% and 60% respectively. The average increase in threshold for Condition 2 and 2a were much smaller 13% and 17% respectively. The thresholds for Condition 1, 2 and 2a were nearly within one SEM (averaged over the observers and three conditions). On the other hand, the threshold for Condition 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 (Condition 2a and 3a) were within one SEM from the conditions with secondary reflections (Condition 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 has little, if any, effect. This effect persists even when the secondary reflections are switched off, 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 surface.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Threshold +- SEM (over sessions) | | | | |
| Subject | 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 |

**Figure S1:** Two example trials of each condition in supplementary experiment 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 observers in the supplementary experiment. The discrimination 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 affect thresholds (Condition 2a and 3a).

1. Kingdom, F.A., *Lightness, brightness and transparency: a quarter century of new ideas, captivating demonstrations and unrelenting controversy.* Vision Res, 2011. **51**(7): p. 652-73.

2. Adelson, E.H., *Lightness Perception and Lightness Illusions.*, in *The New Cognitive Neurosciences, 2nd edition*, M. Gazzaniga, Editor. 2000, MIT Press: Cambridge, MA. p. 339-351.

3. Materials, A.S.f.T.a., *Standard test method for luminous reflectance factor of acoustical materials by use of integrating-sphere reflectometers.* Renovations of Center for Historic Preservation, 2017 **98(A)**: p. E1477.

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

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

6. Cottaris, N.P., et al., *A computational-observer model of spatial contrast sensitivity: Effects of wave-front-based optics, cone-mosaic structure, and inference engine.* J Vis, 2019. **19**(4): p. 8.

7. Singh, V., et al., *Computational luminance constancy from naturalistic images.* J Vis, 2018. **18**(13): p. 19.

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

9. Heasly, B.S., et al., *RenderToolbox3: MATLAB tools that facilitate physically based stimulus rendering for vision research.* J Vis, 2014. **14**(2).

10. Jakob, W., *Mitsuba renderer.* 2010.

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

12. Prins, N. and F.A.A. Kingdom, *Applying the Model-Comparison Approach to Test Specific Research Hypotheses in Psychophysical Research Using the Palamedes Toolbox.* Frontiers in Psychology, 2018. **9**: p. 1250.

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

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

15. Brainard, D.H., *Color and the Cone Mosaic.* Annu Rev Vis Sci, 2015. **1**: p. 519-546.

16. Commission Internationale de l’e ́clairage, *Colorimetry, second edition (Tech. Rep. No. 15.2)*. 1986: Bureau Central de la CIE.

17. Hecht, S., S. Shlaer, and M.H. Pirenne, *Energy, Quanta, and Vision.* J Gen Physiol, 1942. **25**(6): p. 819-40.