Title: Equivalent noise characterization of human lightness constancy

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

**Abstract:** An important goal for vision is to provide stable perceptual 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 brightness discrimination, we measured how thresholds for discriminating the lightness of a target objects are impacted by variation in the reflectance functions of background objects. 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 lightness. On each trial, two images of the scene were presented in sequence and observers indicated which 250ms interval contained the sphere with higher lightness. 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 reflectance functions associated with the background objects. For low variation, the discrimination thresholds were nearly constant, indicating that the observers internal noise determines the threshold. As the scalar increases, the external noise starts to dominate, resulting in higher discrimination thresholds. 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 variations in terms of the intrinsic difficulty of the task.

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

**Introduction:** Visual perception involves the identification of distal properties of an object from the proximal signal to the visual system. While the distal properties are intrinsic to the object, the proximal signal depends on many object extrinsic factors in the visual scene. The challenge to visual system is to recover a stable correlate of the distal object property under variations in the proximal signal due to the object extrinsic factors. As an example, consider the task of recovering the lightness of an object from the light reflected from the object surface. Lightness is an intrinsic property that depends on the surface reflectance spectrum of the object. It is related to the amount of the light that is reflected from an object’s surface [Cite paper with definition]. On the other hand, 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 [1]. But the relation between the amount of variation in the background to the corresponding variation in the perceived lightness is not quantified. In this work, we empirically establish this relation for variation in the spectra of background objects. To do this, we measure the threshold of discriminating two objects based on their lightness. We study how this threshold changes as we vary the color of objects in the background. The color of background objects is randomized by sampling from a statistical model based on the reflectance spectra of natural surfaces. The threshold is measured as a function of a scalar that multiplies the covariance matrix of the distribution. By varying the scalar from 0 (no variation) to 1 (variation in natural scenes), we quantify the effect of background color variation in terms of the intrinsic difficulty of the lightness discrimination task (at scalar = 0). 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 start to rise are consistent over different observers. Moreover, a computational model that estimates the lightness of an object using a center-surround receptive field captures the essential features of human observers.

**Results:**

1. **Measurement of lightness discrimination thresholds**

Our goal in this work was to study how the variation in the color of background objects affects the lightness discrimination thresholds. The psychophysical task in our experiment is shown in Figure 1. We used a two-alternate forced-choice experiment to measure the thresholds. In each trial of our psychophysical task, human observers viewed two images, a standard image and a comparison image (Figure 1a). The images were shown on a calibrated monitor for 0.25s each one after the other. A 0.25s inter-stimulus interval of a blank screen was presented between the two images (Figure 1b; also See Methods: Experiment Details). The images were 2D computational rendering of a 3D visual scene. Each image contains a gray spherical object at the center. The observer’s task was to report the image in which the spherical object was lighter in color. The surface lightness was defined in terms of the luminous reflectance factor (LRF, [2]). LRF is the ratio of the luminance of a surface under a reference illuminant to the luminance of the reference illuminant itself. We chose the reference illuminant as standard daylight D65. We kept the lightness of the target object in the standard image fixed and varied the lightness of the target object in the comparison image. The order of presentation of the standard and comparison images were chosen in pseudo random order.

We recorded the proportion of times the subjects chose the comparison image to have the lighter target object. The proportion comparison chosen data was measured at 11 values of the target object lightness. Figure 2 shows the psychometric function of 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 Curve). We defined the threshold as the difference between the LRF of the target object at percent comparison chosen 76% and 50%. The LRF threshold was obtained from the cumulative Gaussian fit to the proportion comparison chosen data.

To study the effect of background color on lightness discrimination threshold, we varied the spectrum of the background objects in the images by sampling it from a statistical model based on natural surface reflectance databases (See Methods: Reflectance and Illumination Spectra, [3, 4]). Briefly, the database of natural surface reflectance functions was projected along their principal component analysis eigenvectors with the largest six eigenvalues. These eigenvalues captured more than 90% of the variance in the databases. The resulting distribution was approximated as multi-normal distribution. The reflectance spectra of background objects were sampled from this multi-normal distribution. The variability of the samples was controlled by multiplying the co-variance matrix of the distribution by a scalar. We generated images from six logarithmically spaced values of the covariance scalar. Figure 3 shows samples of images used in our psychophysical task. The target object in each image has the same lightness. The color of background objects has been sampled randomly. For the images in the same column, the color of the background objects were sampled at a fixed variance of the distribution. The variance in the samples increases from left to right along the row. The images in the leftmost column have no variation in the background (scale factor = 0), while the images in the rightmost column corresponds to the variation observed in natural scene (scale factor = 1). Discrimination thresholds were measured separately for each of the six values of the covariance scalar.

1. **Human lightness discrimination thresholds increase with the variation in color of background objects**

Figure 4 shows the variation in discrimination thresholds with the amount of variability in the color of the background objects. We plot the log threshold squared vs the log of the covariance scalar of the distribution. The thresholds are averaged over the four subjects. For low values of the covariance scalar, the log squared threshold is nearly a constant and then rises approximately linearly. The variation can be approximated using a double linear function (See Methods: Double Linear Model Fit). At low values of the covariance scalar (low variation in the color of background objects), the variation in the response is dominated by the observer’s internal noise. Thus, the discrimination thresholds remain almost independent of the extrinsic variation. In this region, the log squared thresholds v/s log covariance scalar data can be modeled by a line parallel to the x-axis. As the value of the covariance scalar (variation in the color of background objects) increases, the extrinsic variations in the visual signal starts dominating the observer’s internal noise. Empirically we observe a linear relation between log squared thresholds and log covariance scalar. The intersection of the two lines determines the equivalent noise, which is defined as the value of the log covariance at which the threshold starts to increase. The equivalent noise allows us to quantify the effect of task-irrelevant property in terms of the intrinsic variation associated with the task.

Figure 5 shows the threshold variation for the individual subjects. Although there is some variation in the rate of increase of the thresholds, the observers are consistent in the equivalent noise and covariance scalar for threshold elevation.

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: Computational Model; [5]). 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. We then calculate the dot product of a center surround receptive field (RF) with the retinal images. 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 matched 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 (Figure 6). The value of the receptive field in the center and surround were chosen such that the dot product of each channel of the receptive field with a uniform image was zero.

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 was chosen such that the threshold of the computational model matched the threshold of the mean human observer at covariance scalar equal to zero. We calculated the equivalent noise and the rate of threshold increase of the computational observer using the double linear model. These values compare well with the values of the mean human observer (HOW???).

**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**. Experimental design and the primary 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/.

**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 monitors were 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](http://psychtoolbox.org/)] and mgl (<http://justingardner.net/doku.php/mgl/overview>). The response was collected using a gamepad (Logitech F310).

In the experiment, 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 75 cm.

**Monitor Calibration:** The monitor was calibrated using a spectroradiometer (PhotoResearch PR650). The spectral range of the spectroradiometer is 380-780 nm with an 8nm bandwidth and an accuracy of +/-2nm. To calibrate the monitor, we focused the spectroradiometer to a patch on the center of the monitor. The patch was of the size 4.8cm x 4.6 cm. The spectroradiometer was placed at a distance of 75 cm from the monitor. This gives a visual angle of 3.65 degrees x 3.51 degrees for the patch. The light from the patch was focused using a lens of aperture XXX. This allows the measurement of the light from a small part of the screen.

**Subject Recruitment and Exclusion:** The observers were recruited from the University of Pennsylvania and the local Philadelphia community. These observers were screened to have normal visual acuity (20/40 or better) and normal color vision (as tested with pseudo-isochromatic plates, CITE). The observers received a compensation for their participation in the experiment. The observers were further screened to have reliable performance at the task. This screening was performed in the first session for each observer, which was considered as a practice session. At the beginning of the practice session, the observers were first familiarized with the task. For this they performed a familiarization acquisition (See Methods: Experiment Detailsfor the definition of an acquisition). In the familiarization acquisition, the observers responded to 40 trials of the task from images with scale factor 0.00 (10 easy trials, 10 moderate and 20 regular trials). In the easy trials, the observer 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 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 acquisitions for images with scale factor 0.00. At the end of the practice session, we calculated the mean threshold of the observer for the last two acquisitions. The observer was excluded from the task if their mean threshold for the last two acquisitions in the practice session exceeded 0.025.

If the observer met all these criteria, they were included for the rest of the experiment. Every observer 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 this experiment was collected during a period of 4 weeks.

Total 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.

**Stimulus Design:** Our aim was to measure the thresholds of discriminating two objects based on their lightness 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 color of the background objects. The color of background objects 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 covariance.

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. 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.

**Experiment 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 lightness of the target object in the standard image and varied the lightness 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. 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 run 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 retained 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 with 3 acquisitions. In each acquisition, we randomly selected the images on the trials from the previously generated image databases. The first five trials of each acquisition were moderate trials (as defined above) to acclimatize the observer to the experimental conditions. The response of 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 sequence of 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 quit the experiment anytime during the acquisition. If an observer quit an acquisition, the data for that acquisition was not saved. No observer quit any acquisition. One observer quit at the beginning of a session due to tiredness for unrelated reasons. The session was rescheduled.

**Stimulus Presentation:** The images were presented on an LCD display. The monitor was located at a distance of 75 cm from the observer. The size of each image was 2.6 cm x 2.6 cm on the monitor, corresponding to 2-degree by 2-degree visual angle. The target object size on the screen in the 2D images was ~1 degree of visual angle in diameter. Each image was presented for 0.25 s (this was a deviation from the preregistration document which mentions the presentation time as 0.5 s), with an inter-stimulus interval of 0.25 s and inter-trial interval of 0.25 s. 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 0.25 s (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 a software called Virtual World Color Constancy (VWCC) (<https://github.com/BrainardLab/VirtualWorldColorConstancy>). VWCC is written using MATLAB. It uses the Mitsuba graphics renderer to render images. To render an image, we first need to 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 created in Blender, an open-source 3-D modeling and animation package (<https://www.blender.org/>). Next, we assigned reflectance spectrum and spectral power distribution function 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 assigned, 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/), [6]). The images were rendered at 31 wavelengths equally spaced between 400 and 700 nm. 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-degree cone fundamentals (T\_cones\_ss2 in the Psychophysics Toolbox). Then the monitor calibration data and standard methods 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].

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 Procedures:** At the beginning of the first experimental session (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 Curve:** The proportion comparison chosen data was used to obtain the psychometric curve 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 subjects chose the comparison image to be lighter. The proportion comparison chosen data was fit with a cumulative Gaussian function using the Palamedes toolbox [8]. The data was fit to obtain all four parameters of the psychometric curve: 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.

**Double Linear Model Fit:** The double linear model was defined as:

where *T* is the threshold, corresponds to the minimum threshold, is related to the equivalent noise and gives the rate of increase in log squared thresholds with . The model was fit to and data using Matlab inbuilt *fit* function for the parameters .

**Computational Model:** The computational model consisted of two parts. The first was to estimate the response of the early visual system to the image database. This was done as described in [7]. The model incorporated typical optical blurring, axial chromatic aberration [9], and spatial sampling of the long (L), middle (M) and short (S) wavelength- sensitive cones [10]. 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 [11] 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 [12]. 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.

In the second part of the model, a dot product of the cone response images was taken with a center-surround receptive field, one each for the L, M and S cone response image. 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 (v\_c) and the surround was at another value (v\_s). v\_c and v\_s were chosen such the dot product of the RF with a uniform field was zero. A gaussian noise with mean zero and variance proportional to the dot product was finally added to the resulting dot product. The proportionality constant was chosen such that the threshold of the model at covariance scalar zero (see Methods: Reflectance and Illumination Spectra) was equal the threshold of the mean observer.

To get the model threshold, we sampled random standard and comparison images from our dataset. The receptive field response (noise-added dot product) of the model 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. We estimated the threshold at 9 logarithmically spaced values [0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1]. The retinal images and the Matlab function to get the model thresholds are given in the supplementary information.

**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 Captions:**

**Figure 1:** (a) Psychophysics task: 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. The target was spectrally flat (grey). The color of objects in the background changed in each trial and each interval. The order of presentation of the standard and comparison images were randomized. The lightness of the target object was kept fixed in the standard image and it was varied in the comparison image to get the threshold of discrimination. Discrimination thresholds were studied as function of the amount of variation in background objects.

(b) The sequence of a typical trial. The response R\_(N-1) indicates the end of the (N-1)th trial. The Nth trial begins 0.25s after the response (Inter Trial Interval, ITI). The Nth trial consists of two 0.25s stimuli presentation intervals with a 0.25s inter-stimulus interval (ISI) in between. The observer responds by pressing a button on a gamepad after the second stimulus has been removed from the screen. The observer can take as long as they wished before making the response (R\_N). The next trial begins 0.25s after the response.

**Figure 2: Psychometric Curve:** The psychometric function of a typical human subject in a typical acquisition. We recorded the proportion of times the subject chose the target in comparison image to be lighter. The LRF of the target object in the standard image was fixed at 0.4 LRF. The LRF of the target object in the comparison image were chosen from 11 linearly spaced values in the range [0.35, 0.45]. 30 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 lapse rate and guess rate 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 percent comparison chosen equal to 0.7604 and 0.5, as predicted by the cumulative normal fit. This figure shows the data for subject CNSU\_0002 for scale factor 0.00 in the first session.

**Figure 3: Variation in background color:** The reflectance spectra of background objects were chosen from a statistical distribution of natural surfaces. 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 lightness. For each value of the scalar, we generated 1100 images, 100 each at 11 linearly spaced lightness levels in the range [0.35, 0.45] LRF. Discrimination thresholds were measured separately for each value of the covariance scalar presenting these images on a calibrated color monitor.

**Figure 4: Background variation increases lightness discrimination threshold:** Log squared threshold vs log covariance scalar for mean human observer and the computational model. We calculated the thresholds at six values of the covariance scalar for human observers. The black circular markers give the mean threshold of four human observers. The error bars represent standard error of the mean. The data can be approximated by a double linear model (). We define the intersection of the two lines of the double linear fit as the equivalent noise. The model parameters are given in the legend. The blue diamond markers show the threshold of the computational model. The model captures human behavior.

**Figure 5: Threshold of individual human subjects:** Log squared threshold vs log covariance scalar for individual human observers. The model data is the same as Figure 4. The equivalent noise for the subjects are similar, while the rise in thresholds has individual variability.

**Figure 6: Center surround receptive field:** The receptive field of the computational model. Three identical RFs were used for the long (L), middle (M) and short (S) wavelength cones. The center is taken to be positive and the surround negative. The values of the center and surround was chosen such that the dot product of the RF with a uniform field is zero. The model response to a retinal image was calculated at the dot product of the RF with the image.

**Supplementary Experiments:**

**Measurement of human object luminance discrimination thresholds under variation in object background:** This experiment was the precursor of the experiment performed in the main paper. In this experiment, our aim was to measure the thresholds for object lightness discrimination under three specific types of background variations. These 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 lightness 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 varies from image to images (even at the same lightness level of the target object) because of secondary reflection of light coming from the background objects. We also measured the the thresholds without secondary reflections for these two conditions.

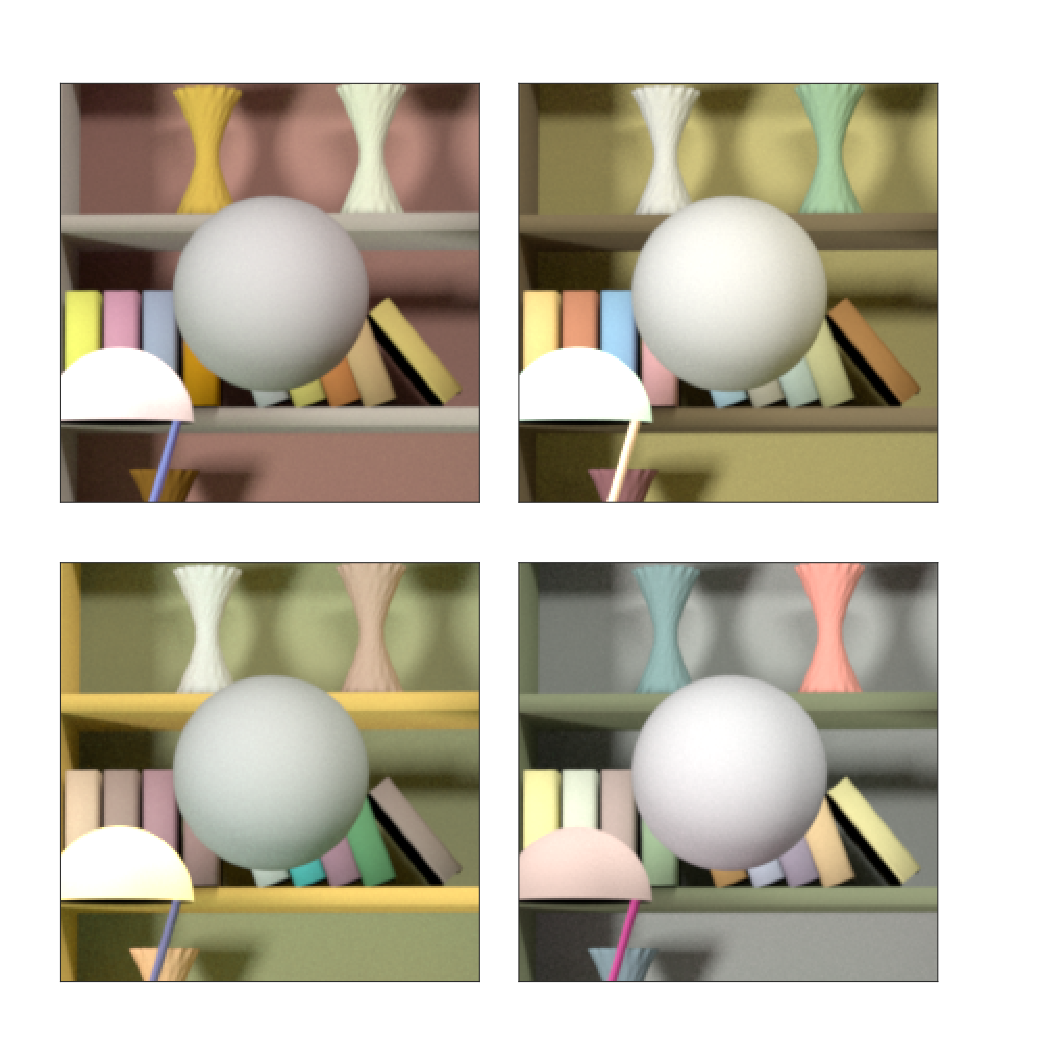
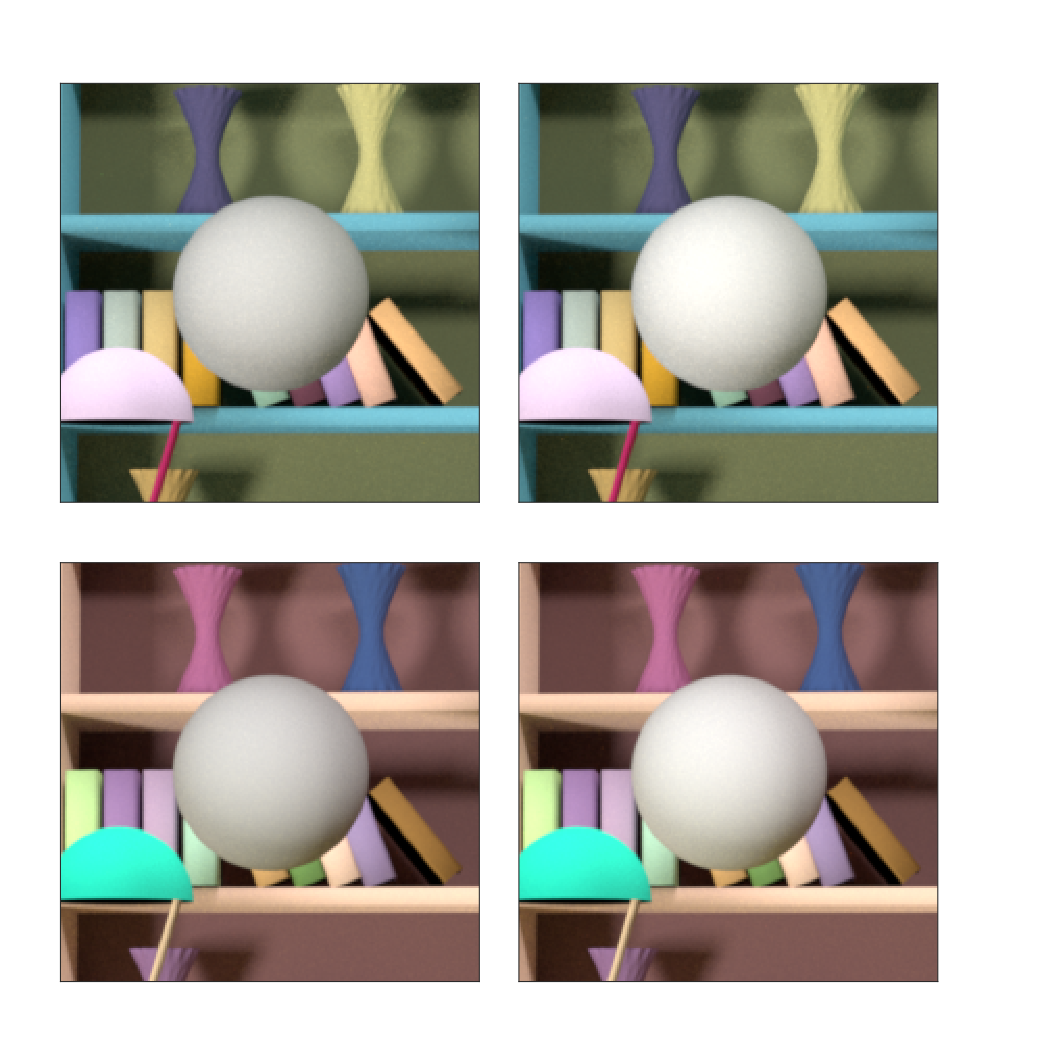


Figure SXXX:Two example stimuli for trials of each condition. For illustration, in this figure we have chosen the stimulus on the left to be at lower LRF. In the experiment, the two images will be presented sequentially in random order at the center of the screen. Conditions 2a and 3a stimuli are similar, but without secondary reflections.

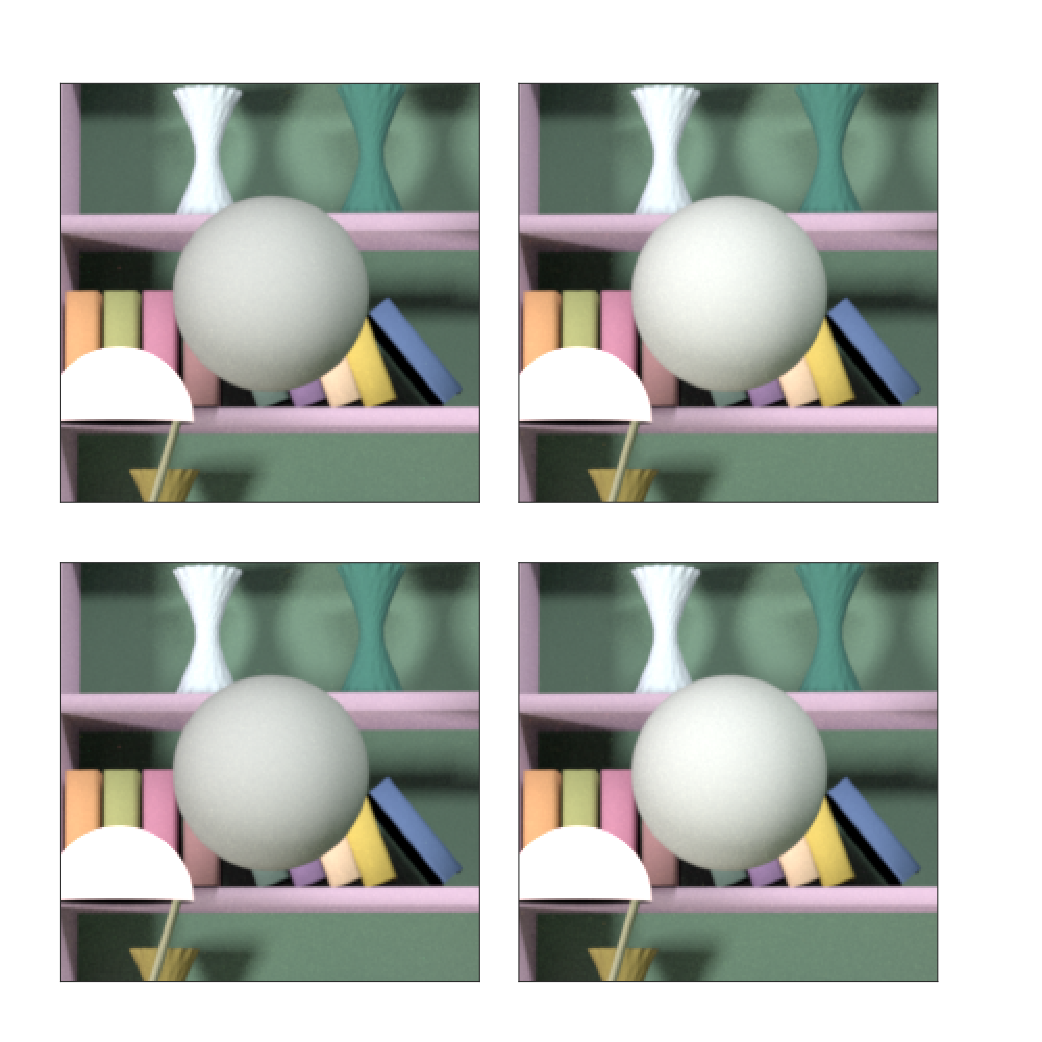
Trial 1

Condition 2

Condition 1

Condition 3

Trial 2



*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.

For this experiment, the reflectance spectra of the background objects were generated with scale factor set to 1.

**Subject Recruitment and Exclusion:** Same as above, except the three acquisition of the practice session was performed with the images in Condition 1.

1. 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.

2. 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.

3. 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.

4. 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.

5. 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.

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

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

8. 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.

9. 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.

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

11. ́clairage, C.I.d.l.e., *Colorimetry, second edition (Tech. Rep. No. 15.2). Bureau Central de la CIE. .* 1986.

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