Response to Referees: "Equivalent noise characterization of human lightness constancy"

We thank the editor and the referees for their comments on our manuscript. We have revised the manuscript as per their suggestions. Below we provide a point-by-point response to the comments. Our comments are in blue. The changes to the manuscript are in green.

Editor Comments:  
  
The two reviews are somewhat divergent, but this is mostly because they focus on different aspects of the paper.  
  
Reviewer #1 is generally more positive, and appreciates the novel methodological advances in the paper. I agree that the paper makes an interesting and potentially important adaptation of methods from spatial vision to lightness constancy.  
  
Reviewer #2 raises concerns about the motivation and presentation of the paper. I think this reviewer raises valid concerns, in two respects. First, the paper may be somewhat opaque to a reader who is not familiar with noise masking models. I understand that this cannot be a tutorial paper, but it may still be possible to give a clearer overview of these models, in order to make the paper accessible to a wider range of readers. Second, the paper could give a stronger motivation of why noise masking models are interesting when applied to lightness perception. What new perspective do these models offer, and what problems might they solve?

*David: I think we can do both of the above. Maybe we add a “Noise Masking” section after the “Introduction” that briefly reviews the logic of noise masking and what they offer when applied to lightness perception/perceptual constancy. We had a little of that in the current intro, but not enough or not sufficiently clear.*

*[[[JDB: Agreed. I think we want to be very clear about how one distinction between the approach here and the traditional noise masking literature is that the dimension that is perturbed and the dimension on which discrimination performance is assessed do not necessarily have to be the same. OLD Literature: Contrast noise is added and contrast discrimination is measured. NEW Idea: Background variation in color (color noise) is added, and LRF discrimination is measured… or similar ]]]*

Reviewer #1:

## Summary  
Equivalent noise paradigm is a psychophysical technique for measuring the internal noise of a human observer, and has been used to characterize the internal noise associated with detection thresholds, including the threshold for letter recognition. In this paper, the authors adapt the paradigm to characterize lightness discrimination thresholds. Using a clever parameterization of known real-world reflectance spectra, authors created scenes with different amounts of "reflectance noise," controlled by the covariance multipler sigma^2: for low noise (low sigma^2), background objects had similar colors, and for high noise (large sigma^2), background colors had widely ranging colors.   
  
Similar to classic findings in contrast detection, the authors found that the square of lightness discrimination thresholds (or rather, the square of the thresholds, T^2) could be separated into a horizontal component, indicating internal noise, and a linearly increasing component, indicating external noise. A model on signal detection theory, with an added constraint for keeping reflectances in [0 1], accounted for most of the data.  
  
## Comments  
The work is excellent. I find it creative that the authors adapted a noise-based method to study lightness discrimination. The manuscript is clearly written and the logic is well laid out.

Thank you for the positive review of our work.  
  
As I am already quite enthusiastic about this work, I bring up the following two discussion points, which the authors already address in the Discussion (Section 3).  
  
- Given the authors' introduction on threshold vs. suprathreshold measures, I was hoping to see how the results from the external noise experiment would connect to a second experiment on suprathreshold perception, possibly by comparing the lightness discrimination thresholds with PSEs.  
  
- I am particularly interested in the possibility of identifying chromatically tuned channels. This could be very exciting future work.

*David: We’re not going to do either of these for this paper, but we can perhaps do a little more work on directions opened up by the current work, which include both of the above.*

Additional minor questions & comments:  
- The section ordering is not typical of a JOV article. Please move Sections 2 (Results), 3 (Discussion) and 4 (Acknowledgements) to the end of the paper.

*Vijay: I have done this.*

Thank you for this suggestion. We have changed the order of the sections, which has improved the paper.

- Why did the authors allow color variations in the background, as opposed to only varying the albedo on neutrally colored objects? Is there any reason to believe that using an achromatic background would change the results? (I would guess no.)

*David: I concede that people seem obsessed with this choice on our part, although to me it seems no less arbitrary than only varying the lightness of grayscale background objects. We can say something about this, probably in the discussion.*

- From what I can tell, the authors do not report any luminance values in the manuscript. It would be useful to get some simple characterizations of the stimuli and the monitor (e.g., mean luminance of the image).

Thank you for pointing this out. We now report the average luminance of the standard image as well as the luminance of the target object for the 11 LRF levels in the Methods section (Section 2.10 Stimulus Presentation). We have added:

The average luminance of the standard image for covariance scalar 0.00 was 47.31 cd/m2. The average luminance of the target object for the 11 LRF levels were [67.04, 67.97, 68.9, 69.81, 70.71, 71.61, 72.49, 73.36, 74.22, 75.07, 75.91] cd/m2.

- What is the advantage of saying LRF instead of albedo?

*David:* *We spent a long time thinking about this, and we wanted to take luminous efficiency into account. Can expand on why in the paper.*

We included LRF instead of albedo to summarize object surface reflectance since LRF incorporates the effects of human luminosity function. We have now explained this in the manuscript.

“We used LRF as it incorporates human luminosity function in quantifying object surface reflectance.”

- Inclusion criterion required 20/40 vision or better. Did the observers wear corrective eyewear?

We have added this information in the methods section. (Section 2.8 Observer Information)

The visual acuities of the observers were: Observer 2, L = 20/30, R = 20/30; Observer 4, L = 20/15, R = 20/20; Observer 8, L = 20/30, R = 20/25; Observer 17, L = 20/20, R = 20/20.

- As the authors note (Section 3, Discussion, p. 13, line 21), the stimuli were rather small: each image subtended 2 deg x 2 deg, and the stimulus sphere took up a 1-deg diameter region within it (Section 5.7). I doubt that the results would be different with a larger stimulus, but it is an unusual choice.

This choice was motivated by the previously reported size of receptive fields in the visual cortex. We have added the following text to the manuscript to explain this point.

“The size of each image was 2.6cm x 2.6cm on the monitor, corresponding to 2° by 2° visual angle. This choice was made because receptive fields in early visual pathways (e.g., retina, primary visual cortex) pool information locally. For instance, the maximum extent of foveal receptive fields in the primary visual cortex is ~ 1° of visual angle (Gattass, Gross, & Sandell, 1981; Gattass, Sousa, & Gross, 1988).”

- How strong is the assumption that real-world reflectance spectra are well-characterized by a multivariate normal?

While the multivariate normal approximation has its limitations, the spectra generated by this model approximate statistics of the underlying natural surface reflectance measurements reasonably. We have now included the following sentences in the manuscript to provide a comparison of the generated samples to natural dataset:

For covariance scalar equal to 1, where the effect would be strongest, the xyY chromaticity of the samples (mean ± std of samples: [x = 0.36 ± 0.05, y = 0.34 ± 0.06, Y = 7.7 ± 3.6]) were comparable to the natural surface dataset ([x = 0.36 ± 0.07, y = 0.34 ± 0.06, Y = 6.2 ± 4.5]).

- Rather than assuming a Gaussian random variable and imposing a realizability constraint, would it make sense to model a noise distribution that is capped in [0 1], such as a variety of beta?

Thank you for this suggestion. The statistical model we have chosen provides a reasonable approximation of the underlying dataset. A comparison of the samples to the natural dataset shows the images are not limited by the model, but the underlying surface reflectance measurements. Additional datasets of natural surface measurements as well as better models would certainly provide renderings that are better representative of natural scenes. We have added the following sentence to the discussion to emphasize this point:

Future refinement of surface reflectance models, through additional surface reflectance measurements and better statistical models, could be used in conjunction with the parameters of the linear receptive field model developed here, without need for new data collection, to refine the estimate of the effect of naturally occurring background variation on object lightness perception.

Reviewer #2:

Summary  
The paper addresses the question how the human visual system establishes stable percepts of objects properties, such as lightness, in face of an unstable proximal input signal, such as luminance. Traditionally the question is experimentally approached by measuring the appearance of target surfaces (supra-threshold judgment) or their discriminability (threshold judgments). Here the authors suggest a different approach inspired by what is known in contrast perception as the equivalent noise paradigm. Instead of (only) varying the reflectance of the target, the authors vary the amount of variation of the reflectances of the background surfaces ("the noise"). They quantify the amount of background variation that is required to significantly affect discrimination performance of the target. They also evaluate two models, an SDT and a receptive field model, with respect to their ability to predict discrimination performance.   
  
This paper provides an original and novel approach to studying lightness, or more generally questions of perceptual constancy. The suggested approach has the potential to tackle questions that cannot be studies with existing paradigms.

Thank you for recognizing the novelty of the approach.

However, in the present form the paper has a number of weaknesses that need to be addressed before one can fully evaluate the potentially new contribution of the suggested approach. Mostly the focus of the paper has not become clear to me and I was missing conceptual explanations. Below I will describe what I mean by that:  
  
1. The equivalent noise idea is mentioned only in the abstract. The transfer between the equivalent noise paradigm in the study of contrast perception and its realization in the present paper is left to the reader. However, this seems to be THE crucial point of the paper so it should be detailed how and if variation in reflectance functions of background objects can be likened to adding noise (white, pink, ...) to a contrast stimulus.

*David: Maybe this can also be addressed in the “Noise Masking” section I suggest above.*

*[[[JDB: I think the reviewer makes a solid point, here. We are relying too much on the reader to connect the dots. Almost always better to be fully explicit.]]]*

2. What is the main focus of the paper? Is the main purpose to introduce a new technique to study constancy phenomena such as lightness or is the purpose to account for a certain type of lightness discrimination data with a particular type of model?

Thank you for this comment. The paper has two goals: to introduce a psychophysical paradigm that to study the effect of task-irrelevant variation to task-relevant variation and to develop a theoretical framework to relate the thresholds task-relevant variation measurements to amount of task-irrelevant variation. We have edited the introduction and added a new section in the manuscript to emphasize the second goal. We have added

We also develop a theoretical framework based on the theory of signal detection to relate thresholds of the task-relevant property to the amount of variation in task-irrelevant property.

Finally, we develop a computational model that quantitatively relates object-extrinsic variations to the object-intrinsic noise in the task.

In the introduction it seems to me that the focus is on establishing a new psychophysical paradigm. However, then I would expect the subsequent data analysis to be more detailed. I would expect to see psychometric functions for each condition and each individual observer (and I would kindly ask the authors to provide those plots). I would expect a reliability estimate of the function/parameter estimates for example by means of confidence intervals.

Thank you for your comment. We have performed further analysis and added two new figures to present our data. We have added Figure 4 which gives the psychometric functions of one of the observers for conditions for each measurement. Additionally, we have provided a supplemental figure with the psychometric functions for each observer for each condition and each measurement. The thresholds were estimated three times for each condition and each observer. We have provided the mean thresholds and the standard error of the mean for each observer. These thresholds are plotted in Figure 5 for the mean observer and in Figure 6 for individual observers. In the figure, we have provided error bars on the measured thresholds represent the confidence intervals as measured by the standard error of the mean. We have also provided the numerical values of the data in these plots (Figures 5 and 6) as a supplementary table (Table S2).

In the results section, the evaluation of the two models is given as much space as the results themselves. That was unexpected because there was no theoretical motivation for the models in the introduction. So after reading the results I was not clear anymore about the main point of the paper, and that vagueness was not removed in the discussion either.

*David: I think the models are deeply tied to the ability to interpret the noise masking data, so here I think the approach is to set that up earlier, perhaps in the “Noise Masking” section.*

3. The focus of the paper is important for another reason. The amount of variation of the background in color or lightness perception is also referred to as articulation or sometimes as contrast range/depth. Usually that has a positive effect on lightness constancy. This interpretation of the experimental variation is quite different from (a low-level) one as "noise". It is still consistent with the data, because one could think of the variations in the background as stabilizing the perceived lightness of the target against fluctuations in its own luminance. That would make the visual system less sensitive to differences in the luminance/reflectance of the target. But it is a different way of talking about the effect of "noise" / variation of background objects.

*David: I think the reviewer is a little confused here. Vijay had suggested we address this point in the first submission, but I took it out because I didn’t think people would be confused. Clearly I was wrong. It is easy enough to explain how ideas of articulation that have been expressed in the constancy literature are expected to have a different effect than the noise we introduce here, but we it’s clear we need to lay that out.*

*[[[JDB: Yup. I think if it is laid out compactly, it should serve to avoid confusion while not constituting too much of a distraction ]]]*  
4. The sequencing of the individual sections is rather unconventional, e.g. Methods are at the end, Results contain methods. I think this is part of the reason why important conceptual explanations are not in the places where they should be. Some important explanations about the models are in the technical details of the Methods sections when they should be introduced on a conceptual level in the Introduction. On the other hand, description of "preregistration" details are not essential to replicate the study and hence should go into supplementary information and so should "observer recruitment details".  
Using terminology such as "acquisition" for a "block of trials" makes the description unnecessarily complicated p17 lines 28-44. There are many places in which essential information needs to be separated from supplementary ones.

Thank you for this comment. We have taken this comment seriously and have restructured the manuscript to move the Methods section before the Results. We have also restructured the subsections of the Methods section as suggested by the referee. We have provided further details of the methods and the methods. We have also added a new section “Noise Models” after the introduction to layout the conceptual ideas of the paper.

To get one estimate of the threshold for one condition, we collected response of 330 trials. These 330 trials were further divided into three sub-sets each of 110 trials, between which the observers were forced to take a minimum 1-minute rest before continuing to the next set. Taking the referee’s suggestion, we have renamed the set of 330 trials as “blocks” and the sub-sets of 110 trials as “sub-blocks”.

In the following I mention some specific points that I came across while reading the paper  
  
Intro  
p2 line 28 "Several theoretical frameworks have been developed, which provide a variety of means for understanding how different cues shape perceptual representations of object reflectance"  
→ please name them, assign references and potentially position yourself

*David: I really hate long introductions and discussions, but there is a world of people who want every paper to not just report what was done but also to serve as a review. That said, we can expand as requested.*  
  
p2 line 33: "Psychophysical methods for measuring discrimination thresholds complement the class of experiments described above." Which ones? Matching experiments (line 21/22)? please name explicitly what you are referring to.

*David: Can clarify.*

Task:  
Why did you provide feedback to participants?

We provided feedback to observers to maximize their performance at the task. We have added the following sentence to explain this point in the manuscript:

Feedback was provided via tones presented after the response to give observers a chance to maximize their performance, an advantage of a task with an objectively correct answer.

Methods  
Is it possible to provide an average luminance of the target for each of the 11 LRF levels?

Thank you for this comment. We have now provided the average luminance of the target object in the manuscript. We have added the following text:

The average luminance of the standard image for covariance scalar 0.00 was 47.31 cd/m2. The average luminance of the target object for the 11 LRF levels were [67.04, 67.97, 68.9, 69.81, 70.71, 71.61, 72.49, 73.36, 74.22, 75.07, 75.91] cd/m2.

Section numbering "5.2." is used twice

Thank you. We have corrected this.  
  
5.8. Image generation and 5.9. Reflectance and Illumination Spectra should be inserted before 5.5 stimulus design. Section 5.9. is difficult to understand without knowing the data set that is referred to in this section. One would need to understand how large was the original space of "natural datasets" (p 19 line 4) and what dimensions it was composed of. Either one needs to go to that level of detail in order to give the reader a chance to understand, or one omits that description and refers to the other paper.

We have reorganized the subsections of the Methods section according to the referee’s suggestion. We have provided further details of the methods to generate reflectance spectra in the revised manuscript. Among other changes, we have added the following text for the information regarding the natural surface reflectance datasets.

We used two datasets of natural surface reflectance functions (Kelly, Gibson, & Nickerson, 1943; Vrhel, Gershon, & Iwan, 1994) containing 632 surface reflectance measurements in total. The Munsell dataset has 462 spectral measurements, each spectrum measured in the wavelength range 380nm to 780nm spaced by 5nm (Kelly, Gibson, & Nickerson, 1943). The Vrhel dataset has 170 spectral measurements, each spectrum measured in the wavelength range 390nm to 730nm spaced by 2nm (Vrhel, Gershon, & Iwan, 1994). We resampled the combined dataset to 31 evenly spaced wavelength between 400nm and 700nm.  
5.11 SDT model  
on page 20 line 1 you write "The fact that we draw stochastically from this ensemble on each trial introduces additional variability into the value of the decision variable z that corresponds to a fixed target LRF. We call this the external variability, and model it as a Gaussian random variable with zero mean and variance σ²\_e".   
  
Here you provide the motivation why you are generating 100 different images for each of the 11 LRF levels of the target. This is important conceptual information that needs to go somewhere in the reasoning about the models. Also, the explanation of the signal detection model can be significantly shortened: you assume the so-introduced external noise to scale with the covariance scalar.

*David: Again, easy to handle. But a little funny that the reviewer asks for a shorter explanation of TSD, in the context of kvetching generally about us being too telegraphic. I’m not sure we want to shorten.*  
  
Results  
page 4 line 27 ff "Briefly, a database of natural surface reflectance functions (Kelly, Gibson, & Nickerson, 1943; Vrhel, Gershon, & Iwan, 1994) was projected along eigenvectors associated with the largest six eigenvalues of the dataset ..."  
page 4 line 33 "The amount of variation in the background was controlled by multiplying the covariance matrix of the multivariate-normal distribution by a scalar."  
→ What was the dimensionality of the dataset?   
→ To me the characterization of the background reflectances is generally not clear. The experimental "scene" consists of a finite number of objects: books, shelf, ... which can be assigned reflectances.   
1. Why do you do the sampling from the database?  
2. Increasing the scalar that is multiplied with the covariance matrix affects the amount of colour variation of the background the objects. So the value of that scalar is the operationalization of background variation (or noise).  
I think the authors should try to make an effort to use the different levels of description of the experimental manipulation where appropriate: technical description (above) vs. phenomenological effect of the technical manipulation (below). Maybe just switch their order of appearance, start with intuition and then provide technical details?

*David: I’m not sure I’m completely following all of these little points, but they all amount to requests for increased clarity and should be easy to address.*

*[[[JDB:*

*A1: We sample from a model of the natural reflectance database to approximate the reflectance spectra of surfaces that occur in the real world.*

*A2: The reviewer is correct in his/her description of the effect of the covariance scalar. We have taken the suggestion to provide both a technical description and plain-English description of the effect of the manipulation.*

*]]]*  
  
  
Please provide psychometric functions as in Figure 2 for each observer and each level of the covariance scalar.

*David: Happy to do so. Let’s think about how to do so in a way that doesn’t distract the more casual reader, though.*

[[[*JDB: Choose a few decent looking ones for inclusion in a main text figure. Maybe from the lowest, middle, and highest background variation conditions to show the progression of slopes. And then shove the rest into a supplement, baby!]]]*

p 4 line 40 "Discrimination thresholds were measured separately for each of the six values of the covariance scalar (Appendix: Table S2)"   
→ See the above point, why is a central result in the Appendix?

We have provided this central result in the Results section of the manuscript. This result has been discussed in the sub-section “Human lightness discrimination thresholds increase with background object reflectance variation”. We have plotted the measured thresholds and the confidence intervals on the measurements in Figures 5 and 6. The values plotted in these figures has been made available to readers as a table in the Appendix. This has been done for completeness while reducing clutter in the main paper. We hope that the referee agrees.

Figure 4  
Mean log threshold squared (averaged across observers, N = 4) is plotted against the log of the covariance scalar

No comments.  
  
Discussion  
- here my problem with the focus of the paper becomes apparent again. Does the modeling complement the experimental paradigm (page 13 line13)? I thought the experimental approach with a conventional threshold comparison between with and without context variation was in itself the NEW approach. If the analysis makes part of the suggested approach that should be made more clear.

*David:We can do this, although I’m not really sure why the reviewer thinks the data can be interpreted without a model. I think all we need to do is more or less say that (“To interpret the data we need a model.”)*

*[[[JDB: I agree with David here.]]]*  
  
"We find that the effect of the external variability introduced by variation of background surface reflectances in naturalistic scenes is within a factor of two of the intrinsic precision of the lightness representation."   
What does that mean?

*David: OK, we’ll have to clarify, since this is the main experimental result reported in the paper.*

[[[*JDB: We do. If I’m reading the sentence correctly, I think its emphasis is potentially confusing. The maximum effect of external variability is partly due to an arbitrary choice on our part about where to cap the maximum value of the covariance scalar. If this is our description of what the ‘equivalent input noise’ is, it is clumsy at best. In either case, a wholesale rewrite is in order to make what we intend to be communicating clear... I just re-read the sentence in context (OPENING OF DISCUSSION). I’m still not clear what it is trying to say. We can/should cut it, perhaps replacing it with another point if we want.* J  
  
Here the results are interpreted from an early vision point of view in terms of early noise and its effect on discrimination (due to inspiration from the contrast literature I suppose).   
  
However, from a lightness constancy point of view one could argue that increased variation/articulation stabilizes perceived lightness against accidental fluctuations, or in other words the equivalence class for the target becomes bigger and hence thresholds increase. What is your take on that view?

*David: This is related to the general confusion about how articulation relates to what we’re doing. I really don’t think there is a strong connection. I will try to work on laying out that viewpoint when we get to it.*