How distinct sources of nuisance variability in natural images and scenes limit human stereopsis

David N. White[1] and Johannes Burge[1,2,3]

[1]Neuroscience Graduate Group, University of Pennsylvania, Philadelphia, PA, 19104

[2]Department of Psychology, University of Pennsylvania, Philadelphia, PA, 19104

[3]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, 19104

March 8, 2024

Keywords — natural stimuli, stereopsis, internal noise, psychophysics, natural image statistics, correspondence problem

1 Significance

Linking properties of the external world, and of sensory stimuli, to how neurons and animals respond has proven an important approach to understanding how the brain works. Much is known about how nervous systems respond to simple stimuli. Less is known about how systems respond during real-world viewing. Using data from two human stereo-depth discrimination experiments, we develop a new approach that reveals how distinct features of natural scenes and images contribute to perceptual performance. Results show that stimulus-by-stimulus variation has highly consistent effects on different people. The approach should have broad application to other animal models and other sensory-perceptual tasks.

2 Abstract

Stimulus variability—a form of nuisance variability—is a primary source of perceptual uncertainty in everyday natural tasks. How do different properties of natural images and scenes contribute to this uncertainty? Using binocular disparity as a model system, we report a systematic investigation of how various forms of natural stimulus variability impact performance in a stereo-depth discrimination task. With stimuli sampled from a stereo-image database of real-world scenes having pixel-by-pixel ground-truth distance data, three human observers completed two closely related double-pass psychophysical experiments. In the two experiments, each human observer responded twice to ten thousand unique trials, in which twenty thousand unique stimuli were presented. New analytical methods reveal, from this data, the specific and nearly dissociable effects of two distinct sources of natural stimulus variability—variation in luminance-contrast patterns and variation in local-depth structure—on discrimination performance, as well as the relative importance of stimulus-driven-variability and internal-noise in determining performance limits. Between-observer analyses show that both stimulus-driven sources of uncertainty are responsible for a large proportion of total variance, have strikingly similar effects on different people, and—surprisingly—make stimulus-by-stimulus responses more predictable (not less). The consistency across observers raises the intriguing prospect that image-computable models can make reasonably accurate performance predictions in natural viewing. Overall, the findings provide a rich picture of stimulus factors that contribute to human perceptual performance in natural scenes. The approach should have broad application to other animal models and other sensory-perceptual tasks with natural or naturalistic stimuli.

3 Introduction

An ultimate goal for perception science is to understand and predict how perceptual systems work in the real world. One approach to achieving this goal is to probe the system with naturalistic stimuli—stimuli that are derived from the natural environment, or bear substantial similarities to them. By examining how stimulus variation characteristic of real-world scenes affects stereo-depth discrimination, we show that performance patterns are similar across different humans, and we partition the effects of distinct stimulus and scene factors on performance—with some surprising results. Further, natural-stimulus variation causes a high degree of stimulus-by-stimulus consistency across observers, consistency that, in principle, could be used to develop and constrain future image-computable models of human perceptual performance.

There is a long tradition of investigating visual performance in human and animal models using simple stimuli and simple tasks. Recent years have been marked by the realization that simple stimuli and tasks may be insufficiently complex to understand how vision works in the real world. When using natural or naturalistic stimuli, however, there are many difficulties in maintaining the rigor and interpretability that has characterized classic research. One important source of difficulty is the sheer number of factors that inject variability into natural retinal images. Some of these factors depend on the environment: the textural patterns on surfaces, the 3D structure of those surfaces, and how the objects that own those surfaces are arranged in 3D space. Other factors are due to the organism and its relationship to the environment, including the optical state of the eyes and the posture and movements of the eyes, head, and body relative to objects in the scene. All of these factors combine to generate many different retinal images, all of which are associated with a particular value of a distal property (e.g. depth) of interest. Such natural-stimulus variability—a form of "nuisance stimulus variability"—impacts neural response (Baddeley, 1997; Felsen & Dan, 2005; A. Iyer & Burge, 2019), and is an important reason that estimation and discrimination of behaviorally-relevant latent variables (e.g. depth, size, 3D orientation) is difficult. In order to perform well, perceptual systems must select for proximal stimulus features that provide information about the latent variable of interest, while generalizing across (i.e. maintaining invariance to) stimulus variation that is not useful to the task. In natural viewing, the computations run by the vision system should minimize, to the maximum possible extent, the degree to which natural-stimulus variability causes variability in human estimation and discrimination in each critical task (Geisler, 1989; Burge & Geisler, 2011; Burge & Geisler, 2012; Burge & Geisler, 2014; Burge & Geisler, 2015; Sebastian et al., 2015, 2017; Burge, 2020; Chin & Burge, 2020).

Here, using binocular disparity as a model system, we report a systematic investigation of how various

forms of natural-stimulus variability impact performance in a depth discrimination task. To approximate natural-stimulus variation, thousands of stimuli were sourced from a natural stereo-image database with co-registered laser-based range data at each pixel using constrained sampling techniques. The sampled stimuli were used to probe human depth-from-disparity discrimination and to determine distinct properties of natural scenes that place limits on human performance. With appropriate experimental designs and data analysis methods, the natural (random) variation across the uncontrolled aspects of the stimuli in each condition provides one with the ability to determine the limits that distinct types of nuisance stimulus variability place on performance.

Two experiments were conducted using the double-pass psychophysical paradigm (Chin & Burge, 2020; Burgess & Colborne, 1988; J. Gold et al., 1999; Neri & Levi, 2006). In contrast to typical 2AFC forced-designs, in which hundreds of responses are collected for each unique stimulus (or trial), double-pass experiments collect two responses for each of two presentations of hundreds of unique stimuli (or trials) in each condition. The conditions of the experiments were defined by different fixation disparities and levels of local-depth variation. These aspects of the stimuli were parametrically manipulated and tightly controlled. Other aspects of the stimuli—luminance-contrast patterns and local-depth structure—were allowed to vary randomly (as they do in natural viewing). We develop new analytical methods that allow us to infer, from the double-pass data, i) the relative importance of natural-stimulus variability and internal noise in limiting performance, and ii) the specific impact that distinct sources of natural-stimulus variability—luminance pattern variability and local depth variability—have on performance.

Several key findings emerge. First, we replicate a performance pattern from the classic literature: discrimination thresholds increase exponentially as targets move farther in depth from fixation. Second, we show that performance limits are increasingly attributable to stimulus variability (rather than internal noise) as the stimuli used to probe performance have more local-depth variability. Third, we show that two distinct types of natural-stimulus variability—luminance-pattern variation and local-depth variation—have distinct and largely separable effects on human performance. Fourth, we find that as stimulus variation becomes more severe, the absolute impact of that stimulus-by-stimulus variation on performance becomes more severe and also becomes more uniform across human observers.

4 Materials and Methods

4.1 Human Observers

All observers had normal or corrected-to-normal acuity. Two of the observers were authors, and the third was naive to the purpose of the study. All observers provided informed written consent in accordance with the declaration of Helsinki. The Institutional Review Board at the University of Pennsylvania approved all protocols and experiments.

4.2 Data and Software

Psychophysical experiments were performed in MATLAB 2017a using Psychtoolbox version 3.0.12. Stimulus sampling and data-analyses were also performed in MATLAB 2017a. Data from this study is available upon reasonable request.

4.3 Apparatus

Stimuli were presented on a custom-built four-mirror haploscope. The haploscope displays were two identical VPixx ViewPixx 23.9 inch LED monitors. Displays were 53.3×30.0 cm in size, with 1920×1080 pixel resolution and a native 120 Hz refresh rate. The maximum luminance of each display was 106 cd/m^2. After light loss due to mirror reflections, the effective luminance was 94 cd/m^2. The mean background gray level of the displays was set to 40 cd/m^2. The gamma function was linearized over 8 bits of gray level.

All mirrors in the haploscope were front-surface mirrors, to eliminate secondary reflections. The mirrors most proximal to the observer were housed in mirror cubes with 2.5 cm circular viewports. The viewports were positioned 65 mm apart, a typical human interpupillary distance. The openings of the cubes limited the field of view to approximately 16° of visual angle.

The optical and vergence distances of the displays were set to 1.0 m. This distance was verified both by standard binocular sighting techniques and via laser distance measurement. At this distance, each pixel subtended 1.07 arcmin. A chin and forehead rest stabilized the head of each observer.

4.4 Stimuli

Stereo-image patches (32×32 pixels each for the left- and right-eye patches) were sampled from 98 large stereo-images (1920×1080 pixels) of the natural environment with co-registered laser range data at each pixel (Burge et al., 2016). Sampling procedures are described below. Stimuli subtended 1° of visual angle,

were spatially windowed by a raised cosine function, and were presented dichoptically. When viewed monocularly, the windowing caused the stimulus to fade into the mean luminance surround. When viewed dichoptically—assuming the patch had uncrossed disparity, which it always did in these experiments—the windowing caused the stimulus to appear behind a fuzzy aperture. Uncrossed fixation disparities (i.e. uncrossed disparity pedestals) of appropriate size were introduced at the stereo-patch sampling stage by cropping the patch from its source image, assuming that a virtual pair of eyes was fixating a point along the cyclopean line of sight in front of the sampled scene location (A. V. Iyer & Burge, 2018). This created stereo-pairs that are geometrically identical to the retinal images that would have formed on the eyes of an observer standing in the original scene. The size of the virtual fixation error was set such that the uncrossed disparity would have the desired value when the stereo-patch was viewed in the haploscope rig.

Each stereo-image patch in the dataset was labeled by the amount of local-depth variation in the imaged scene region, as quantified by disparity-contrast. Disparity-contrast is given by the root-mean-squared difference between the vergence demand of the central corresponding point and the vergence demands of the points in the local surround

$$c_\delta = \sqrt{\frac{\sum_{\mathbf{x}} (\mathbf{v}(\mathbf{x}) - \mathbf{v}_0)^2 \ \mathbf{w}(\mathbf{x})}{\sum_{\mathbf{x}} \mathbf{w}(\mathbf{x})}}, \tag{1}$$

where $\mathbf{w}$ is the raised cosine weighting function that windows the image, $\mathbf{x} = \{x, y\}$ is the spatial location of each pixel, $\mathbf{v}_0$ is the vergence angle that is required to fixate the 3D-scene point specified by the center pixels of the left- and right-eye image patches, and $\mathbf{v}(\mathbf{x})$ is the vergence angle required to fixate the scene points corresponding to the other pixels in the patch. Note that the difference in vergence demand $\mathbf{v}(\mathbf{x})$ - $\mathbf{v}_0$ is simply equal to the relative disparity between the center pixel and the other pixels in the patch. The vergence demand at each point in the patch was computed for an observer viewing the stimulus at the viewing distance and direction set by the experimental rig (i.e. 1 meter away, straight-ahead) across patches that were 32x32 pixels in size.

Each stereo-image patch was contrast fixed to the median root-mean-squared (RMS) contrast (i.e. c_{rms}=0.3) in the natural-stimulus dataset. RMS contrast is given by

$$c_{\text{rms}} = \sqrt{\frac{\sum_{\mathbf{x}} \ \mathbf{c}^2(\mathbf{x}) \ \mathbf{w}(\mathbf{x})}{\sum_{\mathbf{x}} \mathbf{w}(\mathbf{x})}}, \tag{2}$$

where $\mathbf{c}$ is a Weber contrast image, $\mathbf{w}$ is the raised cosine weighting function that windows the image, and

$\mathbf{x} = \{x, y\}$ is the location of a given image pixel.

4.4.1 Stimulus sampling

Left- and right-eye image patches from a natural-scene database were sampled (i.e. centered) on corresponding points (Burge et al., 2016). Because the stereo-photographs were of natural scenes, each local patch was characterized by a different luminance pattern and by some amount of local-depth variability (see Fig. 1B). Corresponding points in the image were determined directly from the range data (see A. V. Iyer and Burge, 2018).

Stimuli were sampled with known amounts of fixation disparity (i.e. pedestal disparities), relative to the corresponding points, up to a maximum of 5 arcmin of uncrossed disparity and known amounts of disparity-contrast. Patches were screened to ensure that the disparity variability within the central region of each patch equaled the nominal fixation disparity within a tight tolerance (see below). Note that because depth varies naturally across any given patch, this central region was the only region of the patch that was guaranteed to equal the nominal fixation disparity. Disparity-contrasts were constrained to be either "high" (0.025-0.117 arcmin) or "low" (0.393-1.375 arcmin). To ensure that each stereo-image patch was unique, patches were not allowed to overlap radially in their source images by more than 10 pixels; this level of overlap was rare.

If the viewing geometry (i.e. distance and direction) of stimulus presentation in an experimental rig does not match the viewing geometry during stereo-image patch sampling, the stereo-specified 3D structure of presented stimulus will be distorted relative to the geometry of the original scene (Held & Banks, n.d.). Stereo-image patches were sampled from all distances and directions, but presented patches at a fixed distance and direction (i.e. one meter away, straight-ahead). Hence, the stereo-specified depth structure during presentation was distorted from that in the original 3D scene. It is possible to prevent these distortions, but only at the cost of distorting the left- and right-eye luminance images. We opted to preserve luminance structure rather than the details of the stereo-specified 3D geometry of the original natural scene. Throughout the article, the disparity-contrast values that are used to characterize the stereo-specified depth variation in each stereo-image patch were set by each patch as it was viewed by the participants in the experimental rig.

4.4.2 Stimulus vetting

Before being included in the experimental stimulus set, stereo-image patches underwent a vetting procedure. The vetting procedure had two primary aims.

The first, most fundamental aim was to ensure accurate co-registration between the luminance and range information in the half-images of each patch. Accurate co-registration was critical for all aspects of the experiment, because the values of the independent variables (i.e. disparity and disparity-contrast) are determined directly from the range data. Although inaccurate co-registration was rare, it was present in a non-negligible proportion of patches. In such cases, the luminance data that observers would have used to estimate disparity would have been inconsistent with the range data used by the experimenters to compute the nominal ground truth disparity. Hence, failing to identify and exclude poorly co-registered patches would mar the accuracy of the results. Potential stereo-image patches were manually vetted by viewing each patch in the experimental rig with onscreen disparities that were nominally uncrossed, zero, or crossed with respect to the screen. Patches that did not pass scrutiny (i.e. that had the wrong depth relationship relative to the screen) were discarded from the pool. The manual vetting procedure was conducted until thousands of unique stimulus patches without co-registration problems were obtained.

The second aim of the vetting procedure—which was enforced programmatically—was to ensure that the center of each stereo-image patch was a coherent target for depth estimation (see above). We required that the most central ($1/8^\circ$ of visual angle $\approx$ 4x4 pixel) region of each patch contained neither a substantial change in disparity (i.e. a disparity-contrast greater than 20 arcsec), or a half-occluded region. Pixels containing half-occluded regions were allowed outside of the most central region. Because regions that are half-occluded have undefined disparity, stimuli including a half-occluded region have undefined disparity-contrast. For patches containing half-occlusions, disparity-contrast was computed by excluding pixels corresponding to half-occluded regions of the scene from the calculation. We did not exclude stimuli with half-occlusions from the dataset because they occur commonly in natural viewing (A. V. Iyer & Burge, 2018).

4.4.3 Stimulus flattening

From the sampled set of natural stereo-image patches—which contain both natural- luminance-pattern variation and natural-depth variation—we also created a "flattened"—but otherwise matched—dataset of stereo-image patches. To convert patches with natural-depth structure into patches with flat depth structure, either the left- or right-eye half-image patch (chosen by random) was replaced by a duplicate of the remaining right- or left-eye half-image patch. This procedure ensured that there is essentially zero-

disparity variation across the patch, such that the disparity pattern specifies a fronto-parallel plane.

4.5 Procedure

Stimuli were presented at the center of a fixation cross-hairs reticle. The reticle was positioned in the center of a circular, 4° diameter, mean-luminance gray area. The circular area was surrounded by a mean-luminance 1/f noise field. The reticle itself consisted of a 2° diameter circle punctuated by hairs jutting outwards at the cardinal and ordinal directions. Hairs were 1° in length and 4.2 arcmin in thickness.

Stimuli were presented using a two-interval forced choice (2IFC) procedure. Each interval had a duration of 250 ms. The inter-stimulus interval was also 250 ms. In one interval of each trial, a stimulus with a standard disparity was presented. In the other interval, a stimulus with a comparison disparity was presented. The order in which the standard or comparison stimulus was presented was randomized.

The task was to report, with a key press, whether the stimulus in the second interval appeared to be nearer or farther than the stimulus in the first interval. Feedback was provided after each response: a high frequency tone indicated a correct response; a low frequency tone indicated an incorrect response.

Psychometric data was collected in a fully-crossed design with disparity pedestal and disparity-contrast as the independent variables. For each combination of disparity pedestal and disparity-contrast, the method of constant stimuli was used for stimulus presentation. Disparity pedestals were defined by one of five standard disparities: $\delta_{std} = [-11.25, -9.38, -7.5, -5.63, -3.75]$ arcmin. Five equally spaced comparison disparities were paired with each standard. Disparity-contrast levels were defined as $\delta_C = [0.025—0.117, 0.393—1.375]$ arcmin, which were labeled "low" and "high" disparity-contrasts respectively. Stimuli in the low disparity-contrast conditions were just-noticeably non-flat to observers. Stimuli in the high disparity-contrast conditions appeared quite noticeably non-uniform in depth. The high disparity-contrast condition contained stimuli that were easily fusible in most cases.

The comparison disparity pedestals and disparity-contrast levels were chosen based on pilot data. Comparison disparities were chosen with the aim that proportion comparison chosen would be approximately 10% on the low-end and 90% on the high-end across the low disparity-contrast condition. Data with 0% and 100% comparison chosen provides no useful information for estimating decision-variable correlation (see subsection "Partitioning the variability of the decision variable" below). Before collecting the data, each observer completed practice sessions to ensure that discrimination performance was stable.

To simulate the stimulus variability that occurs in natural-viewing conditions, a unique natural stereo-image patch was presented on each interval of each trial. This feature of the experimental design represents

a departure from more standard experimental designs, in which either the same stimulus is presented many times each or stimulus differences (e.g. different random dot stereograms) are considered unimportant and not analyzed.

Experiments were conducted using a double-pass experimental paradigm. In double-pass experiments, observers respond to the exact same set of unique trials two times each. Double-pass experiments enable one to determine the relative importance of factors that are repeatable across trials (e.g. external stimulus variation), and factors that vary randomly across trials (e.g. internal noise).

Two double-pass experiments were conducted. In one, all stimuli had natural-depth variation. In the other, all stimuli were "flattened" (see sub-subsection "Stimulus flattening" above). Importantly, both double-pass experiments used the same scene-locations (and hence, near-identical luminance contrast patterns). This design feature allowed us to examine the relative importance of luminance-pattern-driven variability and disparity-contrast-driven variability in the decision variable (see subsection "Partitioning the externally-driven component of the decision variable" below).

Over the course of each double-pass experiment, 10,000 unique stimuli were presented in 5000 unique trials of each double-pass experiment. Five hundred trials were collected in each of ten conditions (5 standard disparities $\times 5$ comparison disparities $\times 2$ disparity-contrasts). Data was collected in 100-trial blocks (i.e. twenty repeats per comparison disparity level per block). The order in which the blocks were run was randomized and counterbalanced across conditions. Two double-pass experiments were conducted, for a total of 20,000 trials per observer.

4.6 Psychometric fitting

Cumulative Gaussian functions were fit to the psychometric data in each condition using maximum likelihood methods. Discrimination thresholds were calculated from the fitted functions. The relationship between the sensitivity index d' (i.e. d-prime) and percent the comparison chosen PC in a two-interval two-alternative forced-choice experiment is given by

$$\mathrm{PC} = \Phi\left(\frac{d'}{\sqrt{2}}\right), \tag{3}$$

where Φ is the cumulative normal function, with d' given by

$$d' = \frac{\Delta\delta}{\sigma_T^2}, \tag{4}$$

where $\Delta\delta = \delta_{\text{cmp}} - \delta_{\text{std}}$ is the difference between the comparison and standard disparities (i.e. the mean value of the decision variable), and σ_T^2 is the variance of the underlying decision variable. (In accordance with standard practices, we assume that decision variable variance is constant for all comparison-disparity levels at a given standard-disparity level—that is, pedestal disparity. The psychometric data is consistent with this assumption.)

The discrimination threshold T is set by choosing a criterion d-prime that defines the just-noticeable difference. In a two-interval, two-alternative forced-choice (2AFC) experiment, threshold is given by

$$T = \sqrt{\sigma_T^2 \; d'_{\text{crit}}}, \tag{5}$$

where d'_{crit} is the criterion d-prime. For computational simplicity, we assume a criterion d-prime of 1.0 such that threshold level performance corresponds to the 76% point on the psychometric function. Thresholds are thus given by the change in the disparity required to go from the 50% to the 76% points on the psychometric function.

Discrimination thresholds were computed from data across both passes of the experiment. When fitting psychometric data across one or both double-pass experiments (see below), thresholds were constrained to change log-linearly across disparity pedestals. Under this constraint, discrimination thresholds in the conditions of a double-pass experiment associated with a given disparity-contrast are specified by

$$T = \sigma_T = \exp(m\delta_{std} + b), \tag{6}$$

where δ_{std} is the standard pedestal disparity, m and b are the slope and y-intercept of the line characterizing the log-thresholds. This constraint is consistent the predictions of normative models of disparity discrimination with natural stimuli, previously reported patterns in psychophysical data (Blakemore, 1970), and the log-linear patterns in the current threshold data (see Figs. 4 and 7). The maximum-likelihood estimates of the parameters defining threshold under the constraint were fit across all conditions having a given disparity-contrast. They are given by

$$\hat{m}, \hat{b} = \arg\max_{m,\ b} \sum_s L_s([\exp(m\delta_{\text{std}}^{(s)} + b)]^2), \tag{7}$$

where L_s is the likelihood of the raw response data in the sth condition, under the assumption that percent correct is governed by a cumulative normal function with mean parameter equal to the sth disparity pedestal

$\delta_{\text{std}}^{(s)}$ and variance parameter equal to $[\exp(m\delta_{\text{std}}^{(s)} + b)]^2$. Finally, the variance of decision variable at each pedestal disparity was obtained by plugging these estimated parameters into Section 4.6.

4.7 Modeling the decision variable

The decision variable can be modeled as a difference between disparity estimates from the stimuli presented on each trial

$$D = \hat{\delta}_{cmp} - \hat{\delta}_{std}, \tag{8}$$

where $\hat{\delta}_{std}$ is the estimate from the stimulus with the standard disparity and $\hat{\delta}_{cmp}$ is the estimate from the stimulus with the comparison stimulus. In accordance with signal detection theory, if the value of the decision variable is greater than zero (and if the observer sets the criterion at zero), the observer will select the stimulus with the comparison disparity. If the decision variable is less than zero, the observer will select the stimulus with the standard disparity.

The decision variable can be more granularly modeled as the sum of two independent random variables. The first random variable accounts for stimulus-driven variability (i.e. variance that is due to nuisance stimulus variability), and has its value set by the particular stimulus (or stimuli) that are presented on a given trial. The second random variable accounts for internal noise, and has its value set randomly on each trial. In a double-pass experiment, across the two presentations of a particular unique trial in a double-pass experiment (i.e. the presentation in the first pass and the presentation in the second pass), the value of the decision variables are given by

$$\begin{aligned} D_1 =& V + W_1, \\ D_2 =& V + W_2, \end{aligned} \tag{9}$$

where V is stimulus-driven contribution to the decision variable, W is a sample of internal noise, and the subscripts index on which pass the trial was presented. Across the two passes of the double-pass experiment, the decision variables can be described as a single two-dimensional random variable $\mathbf{D} = [D_1, D_2]^\intercal$.

The stimulus-driven component of the decision variable on a single pass of the experiment $V \sim \mathcal{N}(\delta_{cmp} - \delta_{std},\ \sigma_E^2)$ is modeled as unbiased and normally distributed with stimulus-driven variance σ_E^2. The noise-driven component of the decision variable $W \sim \mathcal{N}(0, \sigma_I^2)$ is modeled as zero-mean and normally distributed with variance σ_I^2. If the external (i.e. stimulus-driven) and internal (i.e. noise-driven) components of the

decision variable are independent, as we assume they are here, the total variance of the decision variable on a given pass is given by the sum of the internal and external components

$$\sigma_T^2 = \sigma_E^2 + \sigma_I^2. \tag{10}$$

4.8 Decision-variable correlation

The correlation of the decision variable across passes is given by the fraction of the total variance that is accounted for by external (i.e. stimulus-driven) factors, the factors that are repeated across passes. Hence, decision-variable correlation is given by

$$\rho = \frac{\sigma_E^2}{\sigma_T^2} = \frac{\sigma_E^2}{\sigma_E^2 + \sigma_I^2}, \tag{11}$$

where σ_E^2 is the component of the decision-variable variance accounted for by external (i.e. stimulus-driven) factors and σ_I^2 is the component of the decision-variable variance accounted for by internal factors (i.e. noise). In order to partition stimulus- and internally-driven sources of variability, we combine estimates of decision-variable correlation and discrimination thresholds (see below). Decision-variable correlation is an integral factor in determining the repeatability of observer responses across passes of a double-pass experiment.

4.8.1 Estimating decision-variable correlation

Decision-variable correlation was estimated via maximum likelihood from the pattern of observer response agreement between passes. The log-likelihood of n-pass response data, under the model of the decision variable, is

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_j N_j \ \log \mathcal{L}_n^j(\boldsymbol{\theta}), \tag{12}$$

where $\boldsymbol{\theta}$ represents the parameter(s) to be estimated, j is a specific pattern of response, N_j represents the number of times a specific pattern of response was measured. For a double-pass experiment ($n = 2$), the set of response patterns are given by the combination of all possible combinations of responses for each pass. The number of patterns of binary responses is $N = 2^n$. For 2IFC experiment, $N = 2^2 = 4$, with patterns of responses $j \in \{[-,-],\ [-,+],\ [+,-],\ [+,+]\}$. Here, we use $+$ to indicate that the comparison was chosen and $-$ indicates the comparison was not chosen.

We model the joint decision variable as a vector drawn from a multivariate normal distribution $\mathbf{D} \sim \mathcal{N}(\mathbf{x}; \mathbf{m}, \Sigma)$ with a mean vector $\mathbf{m}$ and covariance matrix Σ. The likelihood of a particular pattern of response is given by

$$\mathcal{L}_2^j(\boldsymbol{\theta}) = \int\limits_{s^j(c_1,c_2)} \mathcal{N}(\mathbf{x}; \mathbf{m}, \Sigma) \; d\mu(\mathbf{x}), \tag{13}$$

where integration is in respect to probability measure μ and s^j is a subset of the support S. Here, s^j defines the integration limits for a specific pattern of response j and is a function of the decision criterion on each pass $c \in \{c_1, c_2\}$. Specifically, the integration limits for each dimension/pass are determined by the values of response pattern. For a response r_i at pass i, where the comparison is not chosen ($r_i = -$), $\mathrm{P}(D_i < c_i)$ and the integration limits are $[-\infty, \; c_i]$. Likewise, for comparison chosen ($r_i = +$), $\mathrm{P}(D_i \geq c_i)$ with integration limits $[c_i, \infty]$.

It is computationally convenient to estimate decision-variable correlation with a normalized joint decision variable $\mathbf{D}^z = [D_1^z, D_2^z]^\intercal$ such that it has unit variance on each pass. Normalizing the joint decision variable sets the normalized means equal to d'. Normalizing the joint decision variable also confers a practical advantage in converting the covariance matrix into a correlation matrix so that it can be fully characterized by decision-variable correlation.

The normalized mean vector and normalized covariance (i.e. correlation) matrix associated with the normalized joint decision variable are given by $\mathbf{m}^z = \mathrm{M}\mathbf{m}$, and $\Sigma^z = \mathrm{M}\Sigma\mathrm{M}$, where the superscript z indicates a normalized parameter, and Σ^z is the correlation matrix (i.e. the covariance matrix of the normalized joint decision variable). The normalizing matrix is given by $\mathrm{M} = \mathrm{diag}(\frac{1}{\boldsymbol{\sigma}_T})$, where $\boldsymbol{\sigma}_T$ is a vector of the standard deviation of the joint decision variable $\mathbf{D}$ in each pass, and where the diag(.) function converts a vector into a matrix with the vector-values on the diagonal. The correlation matrix is given by

$$\Sigma^z = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{14}$$

Substituting parameters associated with the normalized decision variable into equations, yields mathematically equivalent expressions of the likelihoods:

$$\mathcal{L}_2^j(\boldsymbol{\theta}) = \int\limits_{s^j(c_1^z,c_2^z)} \mathcal{N}(\mathbf{x}^z; \mathbf{m}^z, \Sigma^z) \; d\mu(\mathbf{x}^z). \tag{15}$$

We also assume that the criteria associated with the normalized decision variable on all passes equals zero, which is justified by the data and by the two-interval, two-alternative forced choice design. In the general case, when this assumption is not made, the decision criteria should also be normalized—that is, the normalized criteria are given by $\mathbf{c}^z = \mathrm{M}\mathbf{c}$. Thus, when analyzing double-pass experimental data under the indicated assumptions, decision variable correlation $\boldsymbol{\theta} = \rho$ is the only parameter that needs to be estimated. Specifically, the maximum-likelihood estimate of decision variable correlation is given by

$$\hat{\rho} = \arg\max_{\rho} \sum_{j} N_j \ \log \mathcal{L}_2^j(\rho), \tag{16}$$

where $N_p \in \{N_{--},\ N_{-+},\ N_{+-},\ N_{++}\}$ is the number of each type of response agreement or disagreement, and $\mathcal{L}_p \in \{\mathcal{L}_{--},\ \mathcal{L}_{-+},\ \mathcal{L}_{+-},\ \mathcal{L}_{++}\}$ is the likelihood of the data given an underlying decision variable distribution specified by the decision variable correlation. The likelihoods are given by

$$\begin{aligned}
\mathcal{L}_2^{--} &= \int_{-\infty}^{c_1^z} \int_{-\infty}^{c_2^z} \mathcal{N}\left(\mathbf{x}^z; \mathbf{m}^z, \boldsymbol{\Sigma}^z\right) \ d\mu(x_1, x_2), \\
\mathcal{L}_2^{-+} &= \int_{-\infty}^{c_1^z} \int_{c_2^z}^{\infty} \mathcal{N}\left(\mathbf{x}^z; \mathbf{m}^z, \boldsymbol{\Sigma}^z\right) \ d\mu(x_1, x_2), \\
\mathcal{L}_2^{+-} &= \int_{c_1^z}^{\infty} \int_{-\infty}^{c_2^z} \mathcal{N}\left(\mathbf{x}^z; \mathbf{m}^z, \boldsymbol{\Sigma}^z\right) \ d\mu(x_1, x_2), \\
\mathcal{L}_2^{++} &= \int_{c_1^z}^{\infty} \int_{c_2^z}^{\infty} \mathcal{N}\left(\mathbf{x}^z; \mathbf{m}^z, \boldsymbol{\Sigma}^z\right) \ d\mu(x_2, x_2).
\end{aligned} \tag{17}$$

4.8.2 Determining the variances of the decision-variable components

With an estimate of the total variance of the decision variable and an estimate of decision-variable correlation, one can estimate the variances of the externally- and internally-driven components of the decision variable. Plugging Eq. (5) into Eq. (11) and rearranging yields an estimate of the variance the externally-driven component of the decision variable

$$\hat{\sigma}_E^2 = \hat{\rho}\hat{\sigma}_T^2. \tag{18}$$

Plugging this estimate into Eq. (10) and rearranging gives an expression for the internally-driven component of the decision variable

$$\hat{\sigma}_I^2 = \hat{\sigma}_T^2 - \hat{\sigma}_E^2. \tag{19}$$

This series of analytical steps was performed for the two double-pass experiments that were conducted: one with natural and one with flattened local-depth variation.

4.9 Partitioning the externally-driven component of the decision variable

To estimate the contributions of luminance-pattern- and local-depth-driven (i.e. disparity-contrast-driven) variability to the decision variable, performance was compared across the stimulus sets with natural and flattened local-depth variation. Recall that the flattened stimulus set effectively eliminates local-depth-variability from the natural-stimulus set—because the disparity pattern in each flattened stimulus specifies a fronto-parallel plane—while leaving luminance contrast patterns essentially unaffected. Hence, because the luminance-pattern-driven component should be essentially the same in both stimulus sets, and because the local-depth-driven component is eliminated in one of the two stimulus sets, an appropriate comparison should reveal the impact of each factor.

To compare performance across the flattened and natural-stimulus sets, we simultaneously analyzed all data from both double-pass experiments using a quasi-quadruple-pass analysis (see below).

4.9.1 Expanded decision variables and correlations

Before explaining in detail how to estimate the contribution of two distinct stimulus-driven factors it is necessary to show how the decision variable depends on these factors in each of the two double-pass experiments. The decision variables in the experiments with flattened and natural-stimuli are given, respectively, by

$$V_\dagger = L, \tag{20}$$

$$V_* = L + B, \tag{21}$$

where L and B denote the the luminance-pattern- and local-depth-driven components of the decision variable, respectively, and $\dagger$ and $*$ indicate, respectively, whether the decision variable corresponds to stimuli that have been flattened (2nd double-pass experiment) or have natural-depth profiles (1st double-

pass experiment). (Note that, for the simplicity of mathematical development, we present the equations here in the Methods section in the opposite order from which the experiments were conducted and presented in the Results section).

Plugging these expanded forms for the externally-driven component of the decision variable into Eq. (9) yields expanded expressions for the decision variables in each of the two double-pass experiments

$$\begin{aligned} D_\dagger &= \overbrace{(L \qquad)}^{V_\dagger} + W_\dagger, \\ D_* &= \underbrace{(L+B)}_{V_*} + W_*. \end{aligned} \tag{22}$$

Clearly, the presence or absence of the local-depth-driven component of the decision variable was the only component that differed across the two double-pass experiments.

Decision-variable correlations across passes in the flattened and natural double-pass experiments, in terms of these new variables, are given by

$$\begin{aligned} \rho_{\dagger\dagger} &= \frac{\sigma^2_{E\dagger}}{\sigma^2_{T\dagger}} = \frac{\sigma^2_L}{\sigma^2_L + \sigma^2_{I\dagger}}, \\ \rho_{**} &= \frac{\sigma^2_{E_*}}{\sigma^2_{T*}} = \frac{\sigma^2_L + \sigma^2_B + 2\mathrm{cov}[L,B]}{\sigma^2_L + \sigma^2_B + 2\mathrm{cov}[L,B] + \sigma^2_{I*}}, \end{aligned} \tag{23}$$

where $\sigma^2_{T\dagger}$ and σ^2_{T*} are variabilities of the decision variable, where σ^2_L and σ^2_B are the luminance-pattern and local-depth-driven contributions to response variability, $\sigma^2_{I_\dagger}$ is the internal noise when only luminance-pattern-driven variability is present, σ^2_{I*} is the internal noise when both luminance-pattern- and local-depth-driven variability is present, $\dagger\dagger$ indicates comparisons across between passes in the double-pass experiment with flattened-depth profiles (i.e. the second double-pass experiment), and $**$ indicates comparisons across passes in the double-pass experiment with natural-depth profiles (i.e. the first double-pass experiment). Clearly, there are five unknowns—σ^2_L, σ^2_B, $\mathrm{cov}[L,B]$, $\sigma^2_{I_\dagger}$, and $\sigma^2_{I_*}$—and, including the threshold equations from each of the two double-pass experiments (see Eqs. (5) and (10)), only four equations. However, by computing decision-variable correlation between passes across each of the two double-pass experiments, a fifth equation is obtained. Specifically,

$$\rho_{\dagger *} = \frac{\sigma^2_L + \mathrm{cov}[L,B]}{\sigma_{T_\dagger}\sigma_{T_*}} = \frac{\sigma^2_L + \mathrm{cov}[L,B]}{\sqrt{(\sigma^2_L + \sigma^2_{I_\dagger})}\sqrt{(\sigma^2_L + \sigma^2_B + 2\mathrm{cov}[L,B] + \sigma^2_{I_*})}}, \tag{24}$$

where $\dagger *$ indicates the cross-double-pass-experiment comparisons. Now, with five equations and five un-

knowns, the equations can be solved.

4.9.2 Estimating decision-variable correlation with expanded decision variables

A novel quasi-quadruple-pass analysis was used to simultaneously estimate $\rho_{\dagger\dagger}$, ρ_{**}, and $\rho_{\dagger *}$, the decision-variable correlations across all four passes of the two double-pass experiments. The quasi-quadruple pass analysis is distinguished from an "ordinary" quadruple-pass analysis because, in an ordinary analysis, all four passes present identical trials. Here, only some of the four passes present trials with identical stimuli (e.g. the flattened stimuli were similar but not identical to the stimuli with natural depth variation). The quasi-quadruple pass analysis allows the three distinct decision variable correlations to take on different values. An ordinary analysis does not allow this flexibility.

For a quadruple-pass (whether quasi or not), the likelihood function is obtained takes the form $\mathcal{L}_n(\boldsymbol{\theta}) = \sum_j N_j \ \log \mathcal{L}_n^j(\boldsymbol{\theta})$ from Eq. (12) above, but across sixteen response patterns

$$j \in \left\{ \begin{array}{llllll} [++++], & & & & & \\ [+++-], & [++-+], & [+-++], & [-+++], & & \\ [++--], & [+-+-], & [+--+], & [-++-], & [-+-+], & [--++], \\ [---+], & [--+-], & [-+--], & [+---], & & \\ [----] & & & & & \end{array} \right\}.$$

The individual likelihoods for these response patterns are extended from Eq. (15), such that

$$\mathcal{L}_4^j(\boldsymbol{\theta}) = \int\limits_{s^j(c_1^z, c_2^z, c_3^z, c_4^z)} \mathcal{N}\left(\mathbf{x}^z; \mathbf{m}^z, \Sigma^z\right) \ d\mu(\mathbf{x}^z), \tag{25}$$

with integration limits s^j as described in the text proceeding Eq. (13). Here, an example likelihood is the region in the space in which the decision variable is positive on all four passes, given by

$$\mathcal{L}_4^{++++}(\boldsymbol{\theta}) = \int\limits_{c_1^z}^{\infty} \int\limits_{c_2^z}^{\infty} \int\limits_{c_3^z}^{\infty} \int\limits_{c_4^z}^{\infty} \mathcal{N}\left(\mathbf{x}^z; \mathbf{m}^z, \boldsymbol{\Sigma}^z\right) \ d\mu(x_1^z, x_2^z, x_3^z, x_4^z). \tag{26}$$

Just as with the double-pass analysis described above, it is convenient to normalize the joint decision variable $\mathbf{D}$ in quadruple-pass analyses via application of a normalization matrix $\mathrm{M} = \mathrm{diag}(\frac{1}{\boldsymbol{\sigma}_T})$. In a

quasi-quadruple-pass analysis, the vector $\boldsymbol{\sigma}_T$ of standard deviations is given by

$$\boldsymbol{\sigma}_T = \begin{bmatrix} \sigma_{T\dagger} \\ \sqrt{\sigma_{T\dagger}\sigma_{T*}} \\ \sqrt{\sigma_{T*}\sigma_{T\dagger}} \\ \sigma_{T*} \end{bmatrix}, \tag{27}$$

with resulting in correlation matrix

$$\Sigma^z = \begin{bmatrix} 1 & \rho_{\dagger\dagger} & \rho_{\dagger*} & \rho_{\dagger*} \\ \rho_{\dagger\dagger} & 1 & \rho_{\dagger*} & \rho_{\dagger*} \\ \rho_{\dagger*} & \rho_{\dagger*} & 1 & \rho_{**} \\ \rho_{\dagger*} & \rho_{\dagger*} & \rho_{**} & 1 \end{bmatrix}. \tag{28}$$

With estimates i) of the total variance of the decision variables from the two double-pass experiments (i.e. $\sigma^2_{T_*}$ and $\sigma^2_{T_\dagger}$) which are obtained from the thresholds, and ii) of the three decision-variable correlations between passes within and across the two double-pass experiments (i.e. $\rho_{\dagger\dagger}$, ρ_{**}, and $\rho_{\dagger*}$), the values of the five unknown parameters can be determined.

Estimates of decision-variable correlation are obtained by maximizing the likelihood of the data under the model. Specifically,

$$\hat{\rho}_{\dagger\dagger},\ \hat{\rho}_{\dagger*},\ \hat{\rho}_{**} = \underset{\rho_{\dagger\dagger},\ \rho_{\dagger*},\ \rho_{**}}{\arg\max} \sum_j N_j\ \log \mathcal{L}_4^j(\rho_{\dagger\dagger},\ \rho_{\dagger*},\ \rho_{**}). \tag{29}$$

We show in the next section how to solve for the contributions of the two distinct natural stimulus-driven factors—i.e. L and B—to the variance of the decision variable.

4.9.3 Determining the variability of the stimulus-driven components

We modeled natural-stimulus variability as being due to two distinct factors: luminance-pattern variability L and local-depth-variability B. To obtain maximum-likelihood estimates of the variance of the luminance-pattern-driven component of the decision variable $\hat{\sigma}^2_L$, the variance of the local-depthnzz-driven component $\hat{\sigma}^2_B$, and the covariance between these two components $\hat{\text{cov}}[L, B]$, from the maximum-likelihood estimates of the three decision-variable correlations (see Eq. (29)), we rearranged Eqs. (23) and (24) for the variables in question. Specifically,

$$\hat{\sigma}_L^2 = \hat{\rho}_{\dagger\dagger}\hat{\sigma}_{T_\dagger}^2, \tag{30}$$

$$\hat{\text{cov}}[L, B] = \hat{\rho}_{\dagger *}\hat{\sigma}_{T_\dagger}\hat{\sigma}_{T_*} - \hat{\sigma}_L^2, \tag{31}$$

$$\hat{\sigma}_B^2 = \hat{\rho}_{**}\hat{\sigma}_{T_*}^2 - \hat{\sigma}_L^2 - 2\hat{\text{cov}}[L, B]. \tag{32}$$

The maximum likelihood estimates indicated in Eqs. (30) and (32) are plotted in the main text Figure 10. The maximum-likelihood estimate of the covariance between the two components (31) tended towards zero, and can safely be ignored.

4.9.4 Fitting constraints

Model parameters were fit via the quasi-quadruple-pass analysis under a pair of constraints. The first constraint was that the disparity-discrimination thresholds used in normalization matrix M (see Eq. (27)) were set to values obtained from linearly constrained threshold fits (see Eq. (7)). The second constraint was that the scaled covariance (i.e. correlation) between the luminance-driven and local-depth-driven components of the decision variable was constrained to lie between -1 and 1. In particular,

$$-1 < \frac{\text{cov}[L, B]}{\sigma_L \sigma_B} < 1, \tag{33}$$

where the scaling factor is given by the product the standard deviations of the two stimulus-driven components. Given that most estimates of interaction term were near zero, we re-fit the model with the more stringent constraint that this interaction term equaled zero. Eqs. (23) and (24) make clear that setting the interaction term equal to zero forces the different decision-variable correlations to have more constrained values with respect to one another than they would be constrained to have otherwise. The log-likelihoods of the models with their best-fit parameters were essentially identical, regardless of whether the interaction term was set equal to zero or not. Non-zero values of the interaction term thus carried little explanatory value.

4.10 Between-observers decision-variable correlation

To derive an expression for between-observers decision-variable correlation, the stimulus-driven component of the decision variable is assumed to be the sum of two independent random variables. (Note that this expansion of the stimulus-driven component is not inconsistent the expansion used in Eq. (21) above.) One is a stimulus-driven component that is shared across observers; the other is a stimulus-driven component that is private to each observer. Specifically,

$$\begin{aligned} D_1 &= \overbrace{(S_1 + P_1)}^{V_1} + W_1, \\ D_2 &= \underbrace{(S_2 + P_2)}_{V_2} + W_2, \end{aligned} \tag{34}$$

where S_1 and S_2 are stimulus-driven components that are identically driven by the stimulus across observers (i.e. the components are proportional $S_1 \propto S_2$, or identical up to a scale factor $S_1 = aS_2$), P_1 and P_2 are the stimulus-driven components that are private to (i.e. uncorrelated between) each observer, and W_1 and W_2 are the respective noise-driven components (see Eq. (9)). The total variance of the stimulus-driven component of the decision variable V_i in each subject i is given by $\sigma^2_{Ei} = \sigma^2_{Si} + \sigma^2_{Pi}$, the sum of the variances in the shared and private components. (Note the overloaded subscript notation. Here, subscripts to denote different subjects. Earlier, subscripts to denoted different passes through the experiment.) Between-subjects decision-variable correlation is given by

$$\rho_{12} = \frac{\mathrm{cov}[S_1, S_2]}{\sqrt{\sigma^2_{T1}\sigma^2_{T2}}}, \tag{35}$$

where σ^2_{T1} and σ^2_{T2} are the total variance of the decision variables in each observer. Of course, these variances include the effects of internal noise. To eliminate the impact of internal noise in the two observers, one can divide through by the square-roots of the within-observer decision-variable correlations to obtain the partial correlation

$$\rho_{12\cdot W} = \frac{\rho_{12}}{\sqrt{\rho_{11}\rho_{22}}} = \frac{\mathrm{cov}[S_1, S_2]}{\sqrt{\sigma^2_{E1}\sigma^2_{E2}}}, \tag{36}$$

where σ^2_{E1} and σ^2_{E2} are the variances of the stimulus-driven component of the decision variable for each observer, and ρ_{11} and ρ_{22} are the within-observer decision-variable correlations for each observer. This partial correlation $\rho_{12\cdot W}$ specifies the degree to which the stimulus-driven components in two different

observers are correlated with each other. High levels of this partial correlation indicate that stimulus-driven components of the two observer are highly similar.

4.10.1 Estimating between-observers correlations

Estimation of between-observers decision-variable correlation within a given experiment also utilized the quasi-quadruple pass methodology, with a few small but important differences. The vector of standard deviations that determined the normalizing matrix is given by

$$\boldsymbol{\sigma}_T = \begin{bmatrix} \sigma_{T_1} \\ \sqrt{\sigma_{T_1}\sigma_{T_2}} \\ \sqrt{\sigma_{T_2}\sigma_{T_1}} \\ \sigma_{T_2} \end{bmatrix}, \tag{37}$$

where subscripts 1 and 2 indicate observer identity, rather than the experiment number. The resulting correlation matrix is given by

$$\boldsymbol{\Sigma}^{\boldsymbol{z}} = \begin{bmatrix} 1 & \rho_{11} & \rho_{12} & \rho_{12} \\ \rho_{11} & 1 & \rho_{12} & \rho_{12} \\ \rho_{12} & \rho_{12} & 1 & \rho_{22} \\ \rho_{12} & \rho_{12} & \rho_{22} & 1 \end{bmatrix}. \tag{38}$$

where ρ_{12} is the between-observer decision variable correlation, and ρ_{11} and ρ_{22} are the within-observer decision variable correlations. The maximum likelihood estimates of these parameters was given by

$$\hat{\rho}_{11},\ \hat{\rho}_{12},\ \hat{\rho}_{22} = \underset{\rho_{11},\ \rho_{12},\ \rho_{22}}{\arg\max} \sum_j N_j\ \log \mathcal{L}_4^j(\rho_{11},\ \rho_{12},\ \rho_{21}). \tag{39}$$

In each of the two experiments, all three unique pairings of observers per experiment were analyzed, so that three between-observers decision variable correlations were estimated for each condition of each experiment.

4.10.2 Between-observer fitting constraints

Constraints for quasi-quadruple between-observers analysis were similar to those for with-observer analysis. First, disparity-discrimination thresholds used in normalization matrix M (see Eq. (37)) were set to values obtained from linearly constrained threshold fits (see Eq. (7)). Second, the scaled partial correlation $\rho_{12\cdot W}$

between the luminance-driven and local-depth-driven components of the decision variable was constrained to lie between -1 and 1. In particular,

$$-1 < \frac{\text{cov[S1,S2]}}{\sigma_{E1}\sigma_{E2}} < 1. \tag{40}$$

5 Results

Three observers collected 20,000 trials each across two double-pass experiments, with the aim of determining how different types of natural stimulus variability—namely, variation in luminance-contrast patterns and variation in local-depth variation—limit sensory-perceptual performance in a depth-from-disparity discrimination task. Comparing performance between two appropriately designed double-pass experiments enables one to dissect the limits placed on performance by distinct types of stimulus-driven uncertainty versus the limits imposed by noise.

In each of the two double-pass experiments, psychometric data was collected with stimuli sampled from scene locations with two different levels of local-depth variability. There were ten conditions total in each experiment—five fixation disparities (i.e. disparity pedestals) crossed with the two levels of local-depth variability (i.e. disparity-contrast; see Methods). In the first double-pass experiment, all stimuli contained natural luminance-pattern variation and natural local-depth variation. In the second double-pass experiment, "flattened" versions of the stimuli used in the first experiment were used such that local-depth variation was eliminated while leaving luminance patterns essentially unaffected (see Fig. 1B).

To obtain the stimuli for the experiments, thousands of stereo-image patches were sampled from a published dataset of stereo-photographs of the natural environment with co-registered laser-based distance measurements at each pixel (Burge et al., 2016). Corresponding points were calculated directly from the range data; a subset of corresponding points is shown in one example stereo-image (Fig. 1C). Sampled patches were centered on corresponding points (Fig. 1D), with known amounts of fixation disparity at the central pixel, as enforced by a custom stereo-image sampling procedure (A. V. Iyer & Burge, 2018). To quantify local-depth variability (i.e. disparity-contrast), ground-truth disparities were computed at each pixel directly from the distance measurements. The routines upon which the sampling procedures were built were accurate with precision better than ±5 arcsec (A. V. Iyer & Burge, 2018). Hence, sampling errors are smaller than human stereo-detection thresholds for all but the very most sensitive conditions (Blakemore, 1970; Cormack et al., 1991).

Stimuli were presented using a two-interval, two-alternative forced choice (2AFC) design (Fig. 2A). The

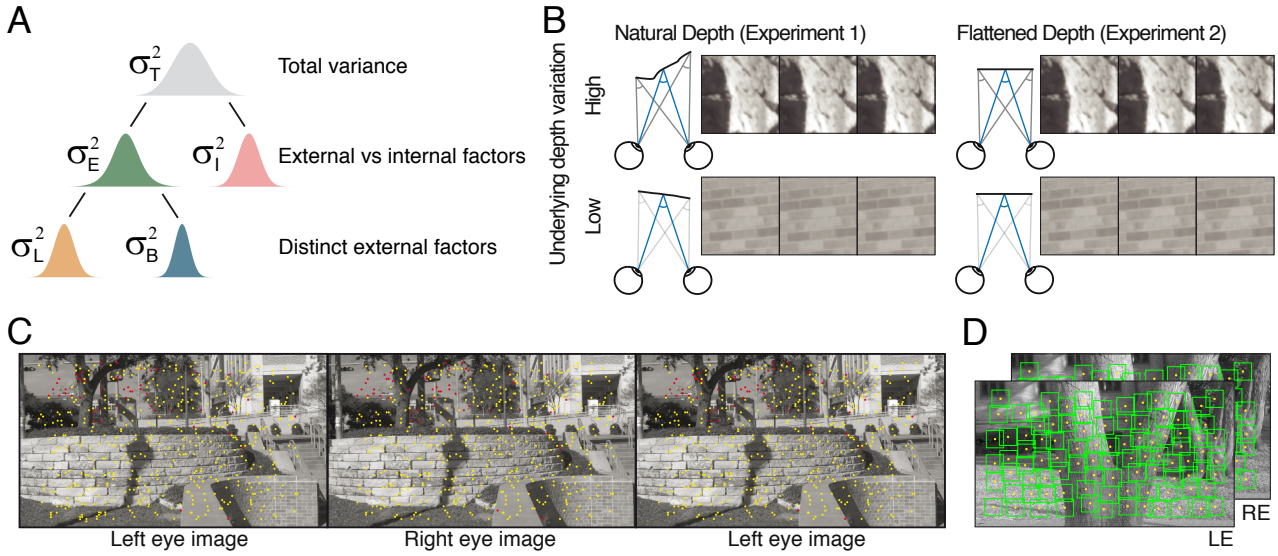

Figure 1. Sources of uncertainty in stereo-depth perception, stereo-image database, and experimental stimuli. **A**. The total variance of the decision variable—the variable that signal-detection-theory posits that perceptual decisions are made on the basis of—is contributed to by at least two distinct sources of uncertainty: external (e.g. stimulus-driven) variability σ_E^2 and internal noise σ_I^2. The stimulus-driven component can be decomposed into distinct external factors: here, luminance-driven variability σ_L^2 and local-depth-driven variability σ_B^2. In natural viewing, luminance-driven variability depends on how luminance-contrast patterns vary across natural stimuli, and depth-driven variability depends on how local-depth structure varies across natural scenes (see *A* and *B*). Traditional psychophysical methods can determine the total variance of the decision variable. Double-pass experiments can partition the total variance into externally- and internally- driven components. The new approach used here can further partition the externally-driven component into distinct external factors. **B**. Two double-pass disparity-discrimination experiments were conducted. Both used images from hundreds of the same natural scene locations. Experiment 1 used stimuli with natural depth profiles (*left*). Local-depth variation, as quantified by disparity-contrast (see Methods Eq. (1)), was either high (*top row*) or low (*bottom row*). Experiment 2 used the same stimuli but with flattened versions of the natural depth profiles (*right*). The flattened stimuli (*right*) had the same luminance profiles as the stimuli in Experiment 1, but had no local-depth variation. **C**. Example natural stereo-image from which natural stimuli were sampled for the experiments, with corresponding points overlaid in yellow. Corresponding points were calculated directly from laser-range-based ground-truth distance data at each pixel. Points in one image without a valid corresponding point in the other (e.g. half-occluded scene regions) are colored red. Divergently-fuse the left two images, or cross-fuse the right two images, to see the scene in stereo-3D. **D**. Another example natural stereo-image with patches that were vetted for inclusion in the experimental stimulus set (*boxes*; see Methods). For purposes of visualization, depicted patches are four times wider (4×4°) than those used in the actual experiments (1×1°).

task was to indicate, with a key-press, which of two natural stereo-image patches, appeared to be farther behind the screen. On each trial, we assume that disparity estimates are obtained for the standard and comparison stimuli: δ_{std} and δ_{cmp}, respectively. Each of these estimates is affected both by properties of the standard and comparison stimuli, and by noise. The decision variable is then obtained by subtracting the standard disparity estimate from the comparison disparity estimates. Distributions of these disparity estimate and decision variable distributions are shown in Figure 2B.

5.1 Decision-variable correlation

The decision variable underlying performance is given by

$$D = V + W, \tag{41}$$

where V is captures the effect of externally-driven, stimulus-based variability and W is a sample of internal noise.

The double-pass experimental design, like a typical (single-pass) experimental design, allows one to estimate the variance of the decision variable. Assuming conditional independence, the total variance of the decision variable is given by

$$\sigma_T^2 = \sigma_E^2 + \sigma_I^2, \tag{42}$$

where σ_E^2 is the variance of the externally-driven component and σ_I^2 is the variance of the internally-driven component. The total variance of the decision variable can be computed directly from the discrimination threshold (Fig. 2B-C). Specifically, for a certain definition of threshold-level performance (i.e. $d' = 1.0$), which we use here, the total variance of the decision variable simply equals the square of the discrimination threshold (i.e. $\sigma_T^2 = T^2$; see Methods Eq. (5)).

The double-pass experimental design, more uniquely, allows one to estimate decision-variable correlation (Fig. 2D-F). decision-variable correlation indicates the degree to which the trial-by-trial values of the decision variable are correlated across passes. It is given by

$$\rho = \frac{\sigma_E^2}{\sigma_T^2}, \tag{43}$$

being equal to the proportion of total variability in the decision variable that is due to factors that are

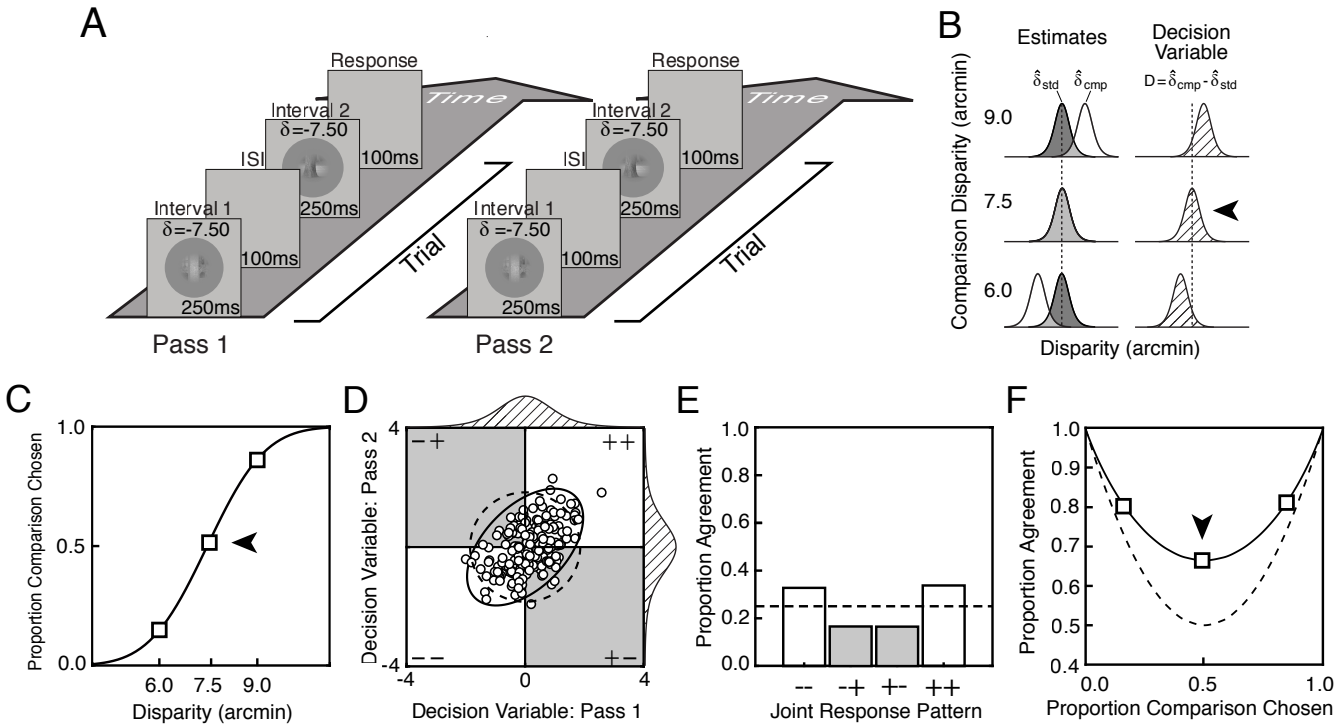

Figure 2. Double-pass experimental design. **A**. Each pass of a double-pass experiment is composed of a large number of unique trials, presented one time each. Each trial is composed of a unique pair of natural stimuli. Trials are identical between passes. The task on each trial is to indicate which of two dichoptically presented stimuli appears to be farther behind the screen. **B**. Standard and comparison disparity estimate distributions for each of three comparison disparity levels (*left*), and corresponding decision variable distributions. Each decision variable distribution is obtained simply by subtracting the standard disparity estimate from the comparison disparity estimate on each trial (*right*). **C**. Psychometric data for stereo-depth discrimination with fitted cumulative Gaussian curve, collapsed across both passes of a double-pass experiment. Thresholds or standard deviation of the decision variable are estimated from the variance parameter of the curve. Psychometric data is binary, indicating whether the comparison stimulus was chosen (+) or not (−). Different decision-variable distributions (*B*) underlie performance at each point on the psychometric function. **D**. Distribution of joint decision variable (*ellipses*) and samples (*dots*) across both passes of a double-pass experiment. Samples in each of the four different quadrants will yield one of four possible joint responses across passes (−−, −+, +−, ++), two of which indicate agreement (++ and −−). Decision-variable correlations larger than 0.0 evince shared sources of response variability across passes. Dashed ellipse shows joint decision-variable distribution if observer responded completely by chance (correlation of zero). **E**. Histograms show the expected proportion of each of the four joint response types for the joint-decision-variable distribution shown in *B*. **F.** Proportion of between-pass agreement as a function of proportion comparison chosen. Solid line shows best fit to the data. Dashed line shows expected agreement if the observer responded completely by chance (correlation of zero).

common across repeated presentations of the same trial (e.g. external stimulus variability) versus those that are not (e.g. internal noise). decision-variable correlation is estimated from the repeatability of observer responses across passes (Fig. 2D-E; see below). On each trial of each pass, the observer reports either that the comparison stimulus appeared farther away than the standard stimulus (+), or that the comparison stimulus appears closer than the standard stimulus (−). Upon completion of both passes, the observer will have made a particular joint response on each unique trial, out of four possible joint responses (−−,−+, +−, ++). When decision-variable correlation equals zero—as it will be when noise is the only source of variability in the decision variable—response agreement is expected to be at chance levels (see Fig. 2D-F, dashed lines). When decision-variable correlation is high—as it will be when external factors (e.g. nuisance stimulus variability) are the dominant source of variance in the decision variable—response agreement is expected to be high.

Decision-variable correlation, like other important quantities in signal detection theory (e.g. d'), must be estimated from a set of binary observer responses (Fig. 2D-F). We computed how repeatable observers' responses were (i.e. how often observer responses agreed) across the repeated presentations of the same stimuli in the first and second passes of the double-pass experiment (see Fig. 2 and Methods). From the level of response agreement in a given condition, we used maximum likelihood techniques to estimate decision-variable correlation across passes.

Decision-variable correlations reflect the *relative* contributions of each individual source of variability in the decision variable (Eq. (41)). A change in decision-variable correlation between conditions could result from an increase in one source of variability, a decrease in the other, or a combination of both. Discrimination thresholds provide an *absolute* measure of the total variance in the decision variable. But they do not indicate the relative contribution of external (e.g. stimulus-driven) and internal (e.g. noise-driven) sources of variability (Eq. (42)). Together, discrimination thresholds and decision-variable correlation can be used to determine the absolute contribution of stimulus-driven and internal-noise-driven sources of variability to the decision variable (see Eqs. (10) and (18)). From estimates of decision-variable correlation (Eq. (43)) and of the total variance of the decision variable (Eq. (42)), the variances of the externally- and internally-driven components of the decision variable can be computed (see Methods, and below).

5.2 Experiment 1: Natural stimuli with natural depth profiles

Figure 3 shows raw data from one individual observer in the first double-pass experiment which used stimuli having natural luminance and natural depth profiles. Psychometric data and function fits showing

proportion comparison chosen are presented in Figure 3A. The slopes of the psychometric functions decrease systematically both as disparity-contrast increases from low to high (top vs. bottom), and as disparity pedestal increases (psychometric functions, left to right). These patterns show that, as the surfaces to be discriminated become more non-uniform in depth (i.e. have higher disparity-contrast), and as they move farther from the fixated distance, discrimination thresholds increase.

Response agreement data and fits for the same observer are shown in Figure 3B. The corresponding estimates of decision-variable correlation in each condition are indicated at the top of each subplot. In all conditions, response agreement is systematically higher than expected under the assumption that decision-variable correlation equals 0.0. Indeed, decision-variable correlation is approximately equal to 0.5, on average across the conditions. Thus, the relative contributions of externally- and internally-driven components to the variance of the decision variable are similar (i.e. $\sigma_E^2 \approx \sigma_I^2$; see Eqs. (42) and (43)). External and internal sources limit performance near-equally. Further, decision-variable correlation is always higher in the high than in the low disparity-contrast conditions (see the inset values of ρ in each subplot). The increase in decision-variable correlation with the level of disparity-contrast entails that the threshold increases are due to more substantial increases in the variance of the stimulus-driven than of the noise-driven component of the decision variable.

Figure 4A shows how stereo-based depth discrimination thresholds change with fixation error (i.e. disparity pedestal) and local-depth variability (i.e. disparity-contrast) for each individual observer, and the observer average. For both disparity-contrast conditions, discrimination thresholds are well-characterized by an exponential function, the signature of which is a straight line on a semi-log plot. This exponential rise in discrimination threshold with pedestal disparity is a classic empirical finding (Blakemore, 1970; Cormack et al., 1991; Badcock & Schor, 1985; McKee et al., 1990; S. B. Stevenson et al., 1992), and is predicted by a normative image-computable model of optimal disparity estimation with natural stereo-images (Burge & Geisler, 2014). The current result provides a psychophysical demonstration that the classic exponential law of human disparity discrimination generalizes to natural stimuli. Because this pattern is robust to the particular stimuli that are used to probe performance, it should be thought of as a feature of how the visual system processes disparity, rather than a consequence of the particular stimuli used to probe performance.

Discrimination thresholds are also higher for stimuli with high disparity-contrast than they are for stimuli with low disparity-contrast. Hence, local-depth variability harms depth discrimination performance. As disparity-contrast increases, thresholds shift vertically in the semi-log plots, such that the two sets of thresholds are parallel to another, indicating that the threshold increases with disparity-contrast are

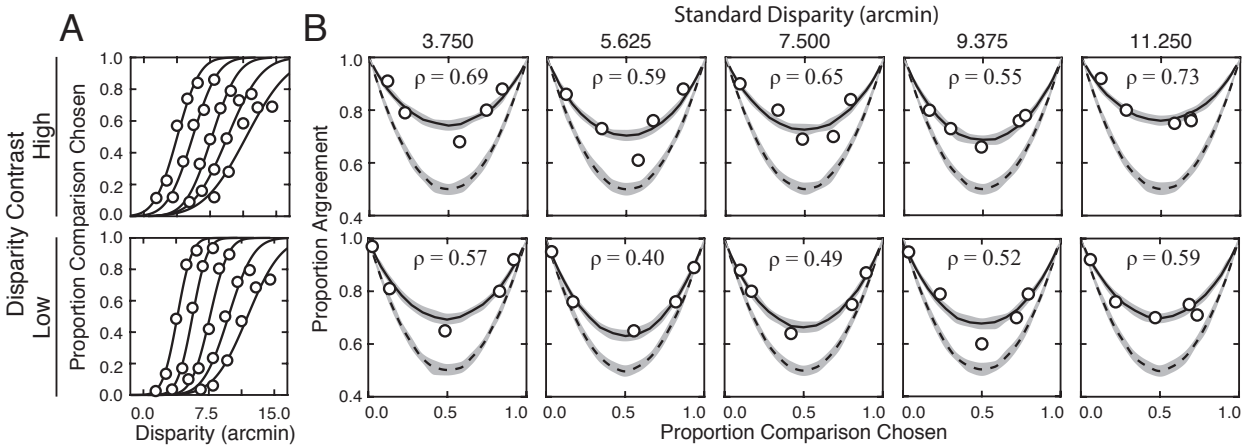

Figure 3. Discrimination thresholds, response agreement, and estimates of decision-variable correlation results for one observer. **A**. Response data (*points*) and psychometric curves for each condition. Thresholds increase systematically with disparity pedestal and with disparity-contrast. **B**. Human response agreement and fitted agreement curves for each condition. Thresholds and decision-variable correlation was used to determine relative impact between sources of performance variability.

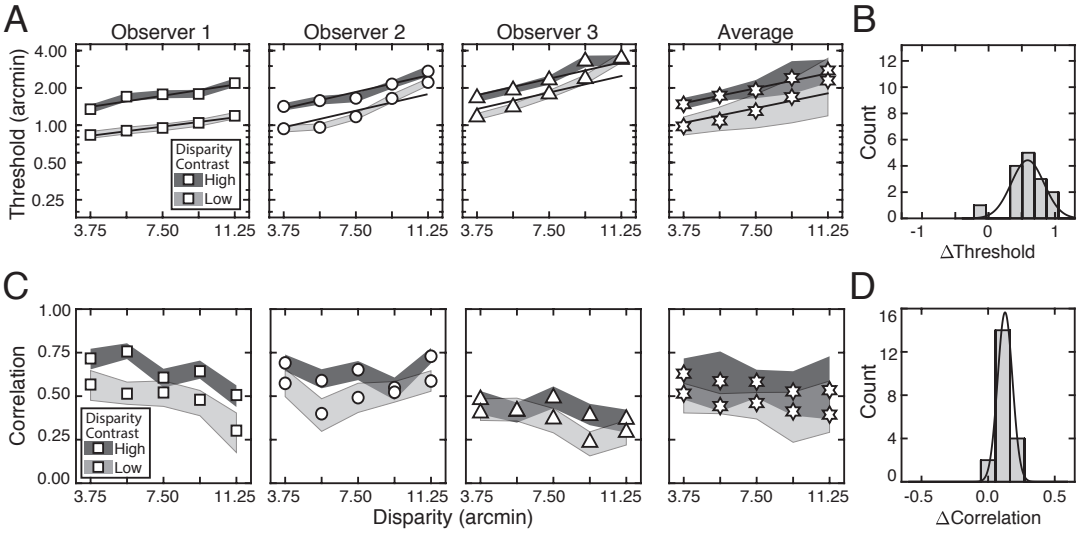

Figure 4. Experiment 1 discrimination thresholds and decision-variable correlations. Stimuli in Experiment 1 stimuli had naturally varying local-depth variation. **A**. Discrimination thresholds as a function of disparity pedestals, for different disparity-contrast levels (*shades*), for each observer and the observer average (*columns*). For individual observers, shaded regions indicate 68% confidence intervals for each condition, generated from 10,000 bootstrapped datasets. For the observer average, shaded regions indicate across-observer standard deviations. Lines represent exponential fits to the data in each disparity-contrast condition (see Methods). Discrimination thresholds are equal to the square-root of the total variance of the decision variable (see Eq. (5)). **B**. Histogram of threshold differences in the high and low disparity-contrast conditions, collapsed across disparity pedestal and individual observers. Curves indicate best-fit normal distributions to the data. **C**. Estimated decision-variable correlation in the same conditions for each observer and the observer average. **D**. Histogram of differences in decision-variable-correlation differences between the high and low disparity-contrast conditions, collapsed across disparity pedestal and individual observers.

multiplicative. Figure 4B visualizes these threshold differences as a histogram, collapsed across all disparity pedestals and observers. Clearly, the histogram of threshold differences is substantially shifted to the right of zero, which confirms that thresholds increase with disparity constraint.

The fact that disparity-contrast degrades discrimination performance should surprise nobody (Banks, 2004; Nienborg et al., 2004; Tyler, 1974; Westheimer, 1979). Increased local-depth variability entails that the left- and right-eye images have more local differences between them. These more pronounced local differences make the stereo-correspondence problem more difficult to solve. The increased difficulty in solving the correspondence problem should, in turn, make stereo-based depth discrimination more difficult. This increase in difficulty is what we observe in our results. However, as we will see, this unsurprising degradation in discrimination performance with disparity-contrast is partly due to a surprising underlying cause (see below).

Decision-variable correlations in each condition for each observer, and for the observer average are shown in Figure 4C. In each and every condition, decision-variable correlation is higher in the high disparity-contrast condition than in the low disparity-contrast condition (Fig. 4D). This consistent pattern of results indicates that as disparity-contrast increases and the task becomes harder, there is an increase in the proportional impact of external, stimulus-driven components on the decision variable—that is, observer responses become more repeatable, not less.

The externally- and internally-driven contributions to threshold were computed from the estimates of decision-variable correlation and the total variance of the decision variables (i.e. discrimination-thresholds squared (see Eqs. (18) and (19)), and are shown in Figure 5. As with the discrimination thresholds (see Fig. 4A)—which reflect the total variance of the decision variable—these individual components also tend to increase exponentially with disparity pedestal (i.e. linearly on semi-log axes; see Fig. 5A). However, disparity-contrast impacts these two components differently. The variance of the external component scales with disparity-contrast (Fig. 5A top row), and substantially so, whereas the variance of the internally-driven component changes more modestly (Fig. 5A bottom row). Thus, the increase in discrimination thresholds with disparity-contrast can be attributed primarily to increases in the variance of the externally-driven (i.e. stimulus-driven) component of the decision variable. The histograms in Figure 5B emphasize this point. They show histograms of the difference in variance between the high and low disparity-contrast conditions in each component, across all observers and disparity pedestals. Clearly, the effect of disparity-contrast on the externally-driven component is more pronounced than the effect on the internally-driven component.

To reiterate, the fact that discrimination thresholds increase with local-depth variability is to be expected

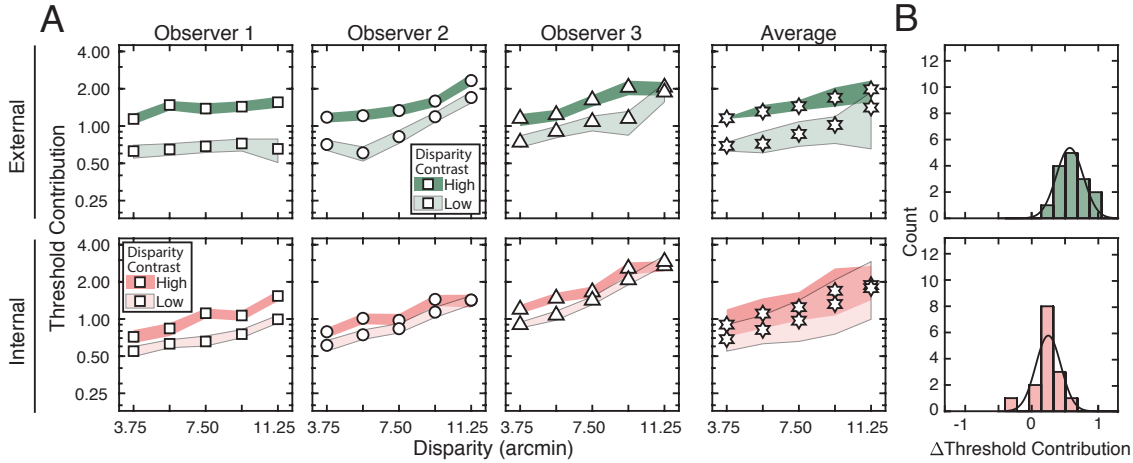

Figure 5. External stimulus-driven and internal noise-driven contributions to thresholds in Experiment 1. **A**. Estimated external stimulus-driven (*top row*) and internal stimulus-driven (*bottom row*) contributions to threshold, at all disparity and disparity-contrast conditions, for each observer and the observer average. For observers, bounds of shaded regions indicate 68% confidence intervals for each condition, generated from 10,000 bootstrapped samples. For the observer average, bounds indicate standard deviations. Threshold contribution reflects the variances σ_E^2 and σ_I^2 of the stimulus-driven and internal noise-driven components of the decision variable, respectively (see Methods). **B**. Histograms of differences between high and low disparity-contrast conditions for both externally- and internally-driven components (*top row* and *bottom row* respectively).

(Banks, 2004). What is unexpected is that a substantial portion of the threshold increases are attributable to factors that make responses more repeatable on successive presentations of the same stimulus. The implication is that, in natural scenes, local-depth variability does not simply make disparity-based depth discrimination noisier, as might be expected if local-depth variability simply made the binocular matching process more unreliable. Rather, the results suggest that local-depth variability biases the observer, stimulus-by-stimulus, to perceive more or less depth in a manner that is repeatable across repeated stimulus presentations. The results therefore imply that, at least in principle, observer errors on each individual stimulus should be predictable. Developing image-computable models that enable stimulus-by-stimulus prediction of depth estimation performance in depth-varying natural scenes is an interesting direction for future work (Burge, 2020).

Another way to investigate the degree to which stimulus-based variability is predictable is to examine between-observer performance similarities. We assessed whether between-observer-threshold variability is more attributable to differences in the effect of external factors (e.g. stimulus-based variability) or internal factors (e.g. noise) across observers. Figure 6 shows how the external, stimulus-based and internal, noise-based contributions to threshold vary across observers relative to the between-observer mean. Between-observer variation in the externally driven-component of the decision variable is substantially smaller than

in the internally-driven component (Fig. 6). The stimulus-driven component of the decision variable is very similar across human observers, and does not contribute substantially to between-observer differences in discrimination threshold. Because the external drive to the decision variable is consistent across observers, it implies that the stimulus-specific computations performed by the human visual system are stable across observers (also see below). Hence, between-observer variability is primarily due to differences in internal noise.

5.3 Experiment 2: Natural stimuli with flattened depth profiles

The second double-pass experiment made use of natural stimuli having "flattened" depth profiles (see Fig. 1B). The luminance profiles of these stimuli are essentially unchanged from those in the first experiment, because they were derived from the exact same scene locations, but the disparity-contrasts of all stimuli were set equal to zero. Thus, in Experiment 2, the nominal "high disparity-contrast" and "low disparity-contrast" stimuli had zero disparity-contrast, even though the luminance profiles were drawn from scene regions originally associated with high and low levels of local-depth variability.

The primary aim of the second double-pass experiment is to make it possible to partition the effects of variation in natural luminance contrast patterns and local-depth variation in limiting stereo-depth discrimination. Doing so requires analyzing the data from both experiments simultaneously. Before turning to this joint analysis of the psychophysical data from both double-pass experiments, we first present the results of the second experiment on their own.

Figure 7 shows discrimination thresholds (i.e. the square-root of the total variance of the decision variable), and decision-variable correlations across all conditions in Experiment 2, for each individual observer

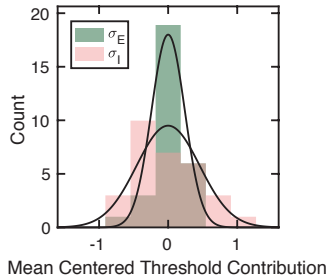

Figure 6. Between-observer variability is primarily attributable to differences in internal noise. Observer-mean subtracted estimates of externally-driven σ_E^2 (green) and internally-driven σ_I^2 (pink) components of the decision variable, histogrammed across conditions. Black lines represent best-fit normal distributions. Across the high and low disparity-contrast conditions, the fraction of between-observer variance explained by the internally-driven component for Experiment 1 was 0.81 ($p = 2.0 \times 10^{-4}$, $F = 0.23$ where F is the test statistic of a two-sample F-test).

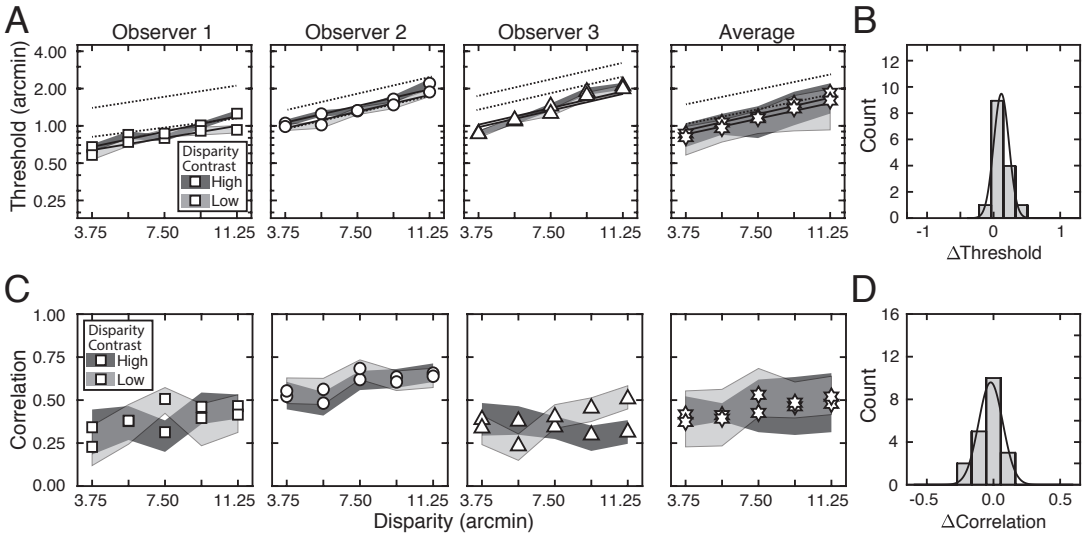

Figure 7. Experiment 2 disparity discrimination thresholds and decision-variable correlation. Experiment 2 stimuli were flattened (i.e. had zero local-depth variability), but otherwise had the same luminance contrast patterns as those in Experiment 1. **A**. Discrimination thresholds as a function of disparity pedestal, for different disparity-contrast levels (*shades*), for each observer and the observer average (*columns*). Unlike in Experiment 1, there is little to no effect of nominal disparity-contrast on threshold. For individual observers, shaded regions indicate 68% confidence intervals for each condition, generated from 10,000 bootstrapped datasets. For the observer average, shaded regions indicate standard deviations. Solid lines represent exponential fits to the data. Dotted lines represent the exponential fits to the threshold data from Experiment 1 (see Fig. 5A). **B**. Histogram of threshold differences in the high and low disparity-contrast conditions, collapsed across disparity pedestal and individual observers. Curves indicate best-fit normal distributions to the data. **C**. Estimated decision-variable correlation in the same conditions for each observer and the observer average. Decision-variable correlations are systematically lower than those in Experiment 1 (see Fig. 4). **D**. Histogram of decision-variable-correlation differences in the high and low disparity-contrast conditions, collapsed across disparity pedestal and individual observers.

and the observer average. There is one marked change in the patterns in the data as compared to the first experiment. Discrimination thresholds (Fig. 7A-B) and decision-variable correlations (Fig. 7C-D) are now largely unaffected by nominal disparity-contrast. There are also consistent decreases in thresholds and decision-variable correlations, as compared to Experiment 1 (see Fig. 4). These results imply that a source of stimulus-driven variance in the decision variable that increases response agreements across repeated stimulus presentations, has been removed from the stimuli.

Analysis of the external (stimulus-driven) and internal (noise-driven) contributions to threshold lead one to the same conclusion: flattening the stimuli removes a stimulus-driven source of variance in the decision variable that is due to local-depth variability (Fig. 8). Neither the external drive to threshold (Fig. 8, top row), nor the internal drive to threshold (Fig. 8, bottom row), is affected by nominal disparity-contrast.

Of course, this change in the pattern of results makes sense. The "high disparity-contrast" and "low disparity-contrast" stimuli in Experiment 2 had been associated with depth varying regions of natural scenes in Experiment 1, but they were flattened for the current experiment. So the result is not unexpected. But it is also not guaranteed. The effect of natural depth variability in bumpier (higher disparity-contrast) scene regions on the decision variable could have been correlated with the effect of natural luminance contrast patterns such that, even with flattened stimuli, the luminance profiles associated with the high disparity-contrast regions of the scene would have generated higher discrimination thresholds. That is, luminance profiles associated with scene locations having greater local-depth variability could themselves have been more difficult to discriminate, even after stimulus-flattening. The current results suggest that this is not the case.

Because of the fact that, in the first double-pass experiment, high disparity-contrast stimuli yielded high levels of externally-driven variance in the decision variable and low disparity-contrast stimuli yielded lower levels of externally-driven variance (see Fig. 5A), the current results strongly imply that a stimulus-driven, and repeatable, source of variability has been removed from the decision variable. The flattened stimuli of the second double-pass experiment also yield the lowest levels of externally-driven variability in the decision variable. Together, these results imply that stimulus flattening removes a distinct source of variability to the decision variable. This idea is tested more rigorously below.

5.4 Partitioning sources of variability in natural stimuli

Here, we show that stimulus-driven variability in the decision variable can be partitioned into separate factors that depend on natural luminance and natural depth structure. These sources of variability—

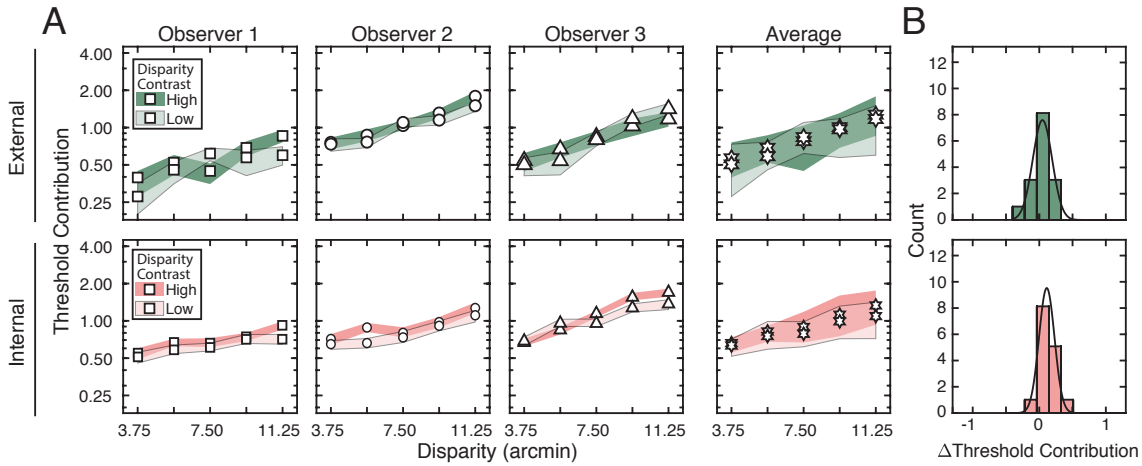

Figure 8. External stimulus-driven and internal noise-driven contributions to thresholds in Experiment 2. **A**. Estimated external stimulus-driven (*top row*) and internal stimulus-driven (*bottom row*) contributions to threshold, at all disparity and disparity-contrast conditions, for each observer and the observer average. For observers, bounds of shaded regions indicate 68% confidence intervals for each condition, generated from 10,000 bootstrapped samples. For the observer average, bounds inidcate across-observer standard-deviations. Threshold contribution reflects the variances σ_E^2 and σ_I^2 of the stimulus-driven and internal noise-driven components of the decision variable, respectively (see Methods). Note that, in comparison to the results of Experiment 1, there is hardly any effect of disparity-contrast on the stimulus-driven contributions to threshold. **B**. Histograms of differences between high and low disparity-contrast conditions for both externally- and internally-driven components (*top row* and *bottom row* respectively).

natural luminance structure and natural depth structure—have distinct and largely separable effects on human performance. To determine the importance of these two factors, and to test whether these factors interact, we compared human performance across the four passes of the two double-pass experiments with flattened and natural stimuli. We refer to this comparative analysis as a quasi-quadruple-pass analysis (see Methods). (Note that typical quadruple-pass experiments—to the extent that quadruple-pass experiments are ever typical—present *exactly* the same stimuli across all four passes. Our experiments presented similar, but not identical, stimuli across the four passes of the two double-pass experiments, hence the "quasi-quadruple-pass" moniker.)

Luminance-contrast pattern variability was essentially the same in both double-pass experiments, and was thus the same across all four passes. However, because the second double-pass experiment used flattened stimuli—which prevents local-depth variability from directly influencing the variance of the decision variable—natural luminance variation is the only remaining stimulus factor that can contribute to the decision variable because natural depth variability has been eliminated. The quasi-quadruple-pass analysis allows one to determine how these two factors combine and/or interact to limit performance.

To understand the reasoning behind the quasi-quadruple-pass analysis, it is useful to write out expanded

expressions for the decision variable (also see Eq. (41) above). The expanded expression for the decision variable is given by

$$D = \overbrace{(L + B)}^{V} + W, \tag{44}$$

where L and B are luminance profile driven and local-depth-variability driven contributions to the decision variable (which sum to the total stimulus-driven contribution V), and W is a sample of internal noise.

In the double-pass experiment with natural luminance and depth profiles (Exp. 1), the expressions for the total variance of the decision variable and for decision-variable correlation across passes, in terms of the variance of these newly articulated components (i.e. L and B in Eq. (44)), are given by

$$\sigma^2_{T_*} = \overbrace{(\underbrace{\sigma^2_L + \sigma^2_B + 2\text{cov}[L, B]}_{\text{unknowns}})}^{\sigma^2_{E_*}} + \sigma^2_{I_*}, \tag{45}$$

$$\rho_{**} = \frac{\sigma^2_{E_*}}{\sigma^2_{T_*}} = \frac{\sigma^2_{E_*}}{\sigma^2_{E_*} + \sigma^2_{I_*}}, \tag{46}$$

where σ^2_L and σ^2_B are the variances of the components driven by luminance profile and local-depth variability, the interaction term $\text{cov}[L, B]$ is the covariance between them (if it exists), $\sigma^2_{E_*}$ is the external (stimulus-driven) variance, and $\sigma^2_{I_*}$ is the variance of internal noise. The external stimulus-driven- and internal noise-driven variances can be solved from the equations for total variance and decision-variable correlation (Eqs. (45) and (46)). But there are not enough equations to separately determine the values of the three unknown factors: the variance σ^2_L of component driven by luminance-pattern variability, the variance σ^2_B of component driven by local-depth variability, and the covariance $\text{cov}[L, B]$ between the luminance and depth driven components. Fortunately, the second double-pass experiment allows one of these unknown factors—the variance of the luminance-driven component of the decision variable—to be determined.

In the second double-pass experiment with natural luminance profiles and flattened depth profiles (Exp. 2), the expanded expression for the decision variable is given by

$$D = \overbrace{(L \qquad)}^{V} + W. \tag{47}$$

Note that the disparity-contrast driven component B that is present in the first experiment does not

appear in Eq. (47), because disparity-contrasts were set equal to zero when the stimuli were flattened. The corresponding expressions for the variance of the decision variable, and decision-variable correlation, are given simply by

$$\sigma^2_{T\dagger} = \overbrace{(\sigma^2_L \qquad\qquad\qquad\qquad)}^{\sigma^2_{E\dagger}} + \sigma^2_{I\dagger}, \tag{48}$$

$$\rho_{\dagger\dagger} = \frac{\sigma^2_{E\dagger}}{\sigma^2_{T\dagger}} = \frac{\sigma^2_{E\dagger}}{\sigma^2_{E\dagger} + \sigma^2_{I\dagger}}, \tag{49}$$

where, again, σ^2_L is the luminance profile driven variance, $\sigma^2_{E\dagger}$ is the external stimulus-driven variance, and $\sigma^2_{I\dagger}$ is the internal-noise-driven variance associated with the flattened stimuli. Just as before, the external and internal variances can be estimated from Equations 48 and 49. Now, the variance of the luminance-pattern-driven component σ^2_L is easily obtained because it exactly equals the variance of the externally-driven component. Also note that in this experiment, because local-depth variability is absent, the variance of the disparity-contrast-driven component is zero. But there are still two remaining unknowns.

Here is where the quasi-quadruple-pass analysis proves useful. By computing decision-variable correlation across passes of the two different double-pass experiments, an additional equation is obtained. Decision-variable correlation between passes across experiments is given by

$$\rho_{\dagger *} = \frac{\sigma^2_L + \mathrm{cov}[L, B]}{\sigma_{T\dagger}\ \sigma_{T*}}. \tag{50}$$

With this expression, we now have the number of equations necessary to determine the unknowns. Using maximum likelihood techniques, we fit all three decision-variable correlations ($\hat{\rho}_{\dagger\dagger}$, $\hat{\rho}_{**}$, and $\hat{\rho}_{\dagger *}$) simultaneously from the data in both experiments with the quasi-quadruple-pass analysis (see Eq. (28)), and then solved algebraically the system of equations specified by Eqs. (45), (46) and (48) to (50) for the unknown parameters. This approach guarantees that shared factors between equations are consistent with one another.

Before proceeding to the main results, we briefly note that we have already estimated decision-variable correlation across passes in the first experiment and in the second experiment—$\hat{\rho}_{**}$ and $\hat{\rho}_{\dagger\dagger}$, respectively—, in each case using data only from the respective experiment in isolation. When carrying out the quasi-quadruple-pass analysis, the estimates of the within-experiment decision-variable correlations (i.e., $\hat{\rho}_{**}$ and $\hat{\rho}_{\dagger\dagger}$) and the variances of the externally- and internally-driven components (i.e., σ^2_E and σ^2_I) are not

guaranteed to be the same as when they are estimated with the data from only one isolated experiment (see Fig. 5 and 8. Reassuringly, however, the estimates from the quasi-quadruple-pass analysis are very similar to those previously estimated. This consistency supports the claim that factors assumed to be common to both experiments are in fact common to both experiments (see Fig. 9). The consistency which these parameters vary across experiments and observers, suggests that each component of the decision variable is indeed driven by the natural-image property—or a tight co-variate of the property—that is said to drive it.

Figure 10 shows the recovered values of the luminance- and depth-driven components of the decision variable—σ_L^2 and σ_B^2, respectively—, and their interaction term $\text{cov}[L, B]$, that were obtained from the quasi-quadruple-pass analysis (see above; also see Methods). The variances of both the luminance-driven and local-depth-driven components clearly increase with disparity pedestal for all conditions and observers. This pattern is similar to the patterns in all previous plots. More interestingly, whereas the luminance-driven component is very nearly unaffected by the level of disparity-contrast (Fig. 10A-B top row), the local-depth-driven component has substantially higher variance with high than for with low disparity-contrast stimuli (Fig. 10A-B bottom row).

These points are emphasized by histograms of the differences in the values of these components in the low and high disparity-contrast conditions. Although the luminance-pattern-driven component is essentially invariant to it (Fig. 10B), the local-depth-driven component changes substantially with disparity-contrast (Fig. 10D). From these results we conclude that the variance of luminance-driven component of the decision variable is a function of pedestal disparity but not disparity-contrast $\sigma_L^2(\delta_{std})$, and that the local-depth-driven component is a function of both factors $\sigma_B^2(\delta_{std}, C_\delta)$, a finding that strongly suggests that the components are not substantively affected by a potential common cause (e.g. local-depth variability). Overall,

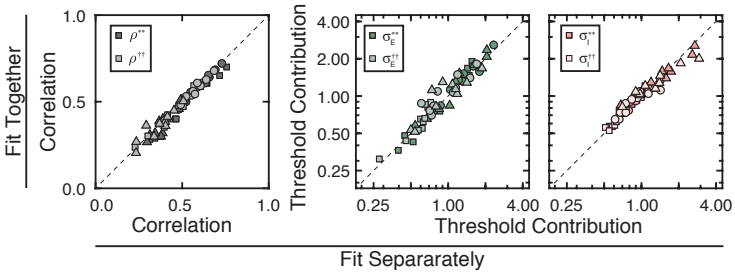

Figure 9. Robustness of fitting methods. Comparison of values obtained from fitting data from Experiments 1 and 2 separately (see Figures 5 and 8) versus together with a quasi-quadruple pass analysis. For decision-variable correlation (*left*), and threshold contributions by stimulus-driven factors (*middle*) and internally-driven factors (*right*), results are consistent regardless of the analytical approach. The consistency of the results indicates the validity and robustness of the quasi-quadruple pass analysis.

these results support the conclusion that natural luminance-pattern variability and natural local-depth variability in real-world scenes have separable effects on stereo-based depth discrimination performance.

Note that the value of the interaction term is near-zero for all conditions (Fig. 10C-D). Refitting the data with a model that fixes the interaction term to zero yields estimates of σ_L^2, σ_B^2, $\sigma_{I_\dagger}^2$, and $\sigma_{I_*}^2$ that are robust to whether the constraint on the interaction term is imposed during fitting; any qualitative description that applies to one set of fitted results applies to the other. Fits with and without the constraint also yield near-identical log-likelihoods. Just as the fitted results are robust to whether data from the two double-pass experiments are fit together (with the quasi-quadruple-pass analysis) or separately (see Fig. 9)), the estimates of whether luminance-pattern- and local-depth-driven sources of variance do not covary with one another.

One might have expected different results. A given scene location gives rise to the luminance patterns in the left- and right-eye images, and to the pattern of binocular disparities between them. So, one might predict that the effect of a given luminance contrast pattern (i.e. photographic content) would be tightly correlated with the effect of the corresponding local-depth variation on disparity discrimination performance, for the simple reason that they have a potential common cause: both are largely determined by the same location in the scene. The current results show that this is not the case. Rather, the results strongly suggest that each of these natural-stimulus-based sources of variability in the decision variable are near-independent of one another.

5.5 Shared stimulus drive between observers

Earlier, we presented data showing that between-observer variability (i.e. threshold differences) was driven more by observer-specific differences in internal noise than by observer-specific differences in stimulus-driven variability (see Fig. 6). We speculated that this result was due to a high degree of similarity between the computations that different humans use to extract useful information from each stimulus for the task. Here, we present data from between-observers decision-variable correlations that bolster the case.

Between-observers decision-variable correlation quantifies the similarity of the decision variable in two different observers across repeated presentations of the same stimuli. If different human observers are using the same computations to estimate and discriminate stereo-defined depth from natural stimuli, stimulus-by-stimulus disparity estimates from one human should be correlated with those from a second—that is, between-observers decision-variable correlation will be substantially larger than zero (assuming internal noise is not too large). On the other hand, if subjects are using quite different computations to process

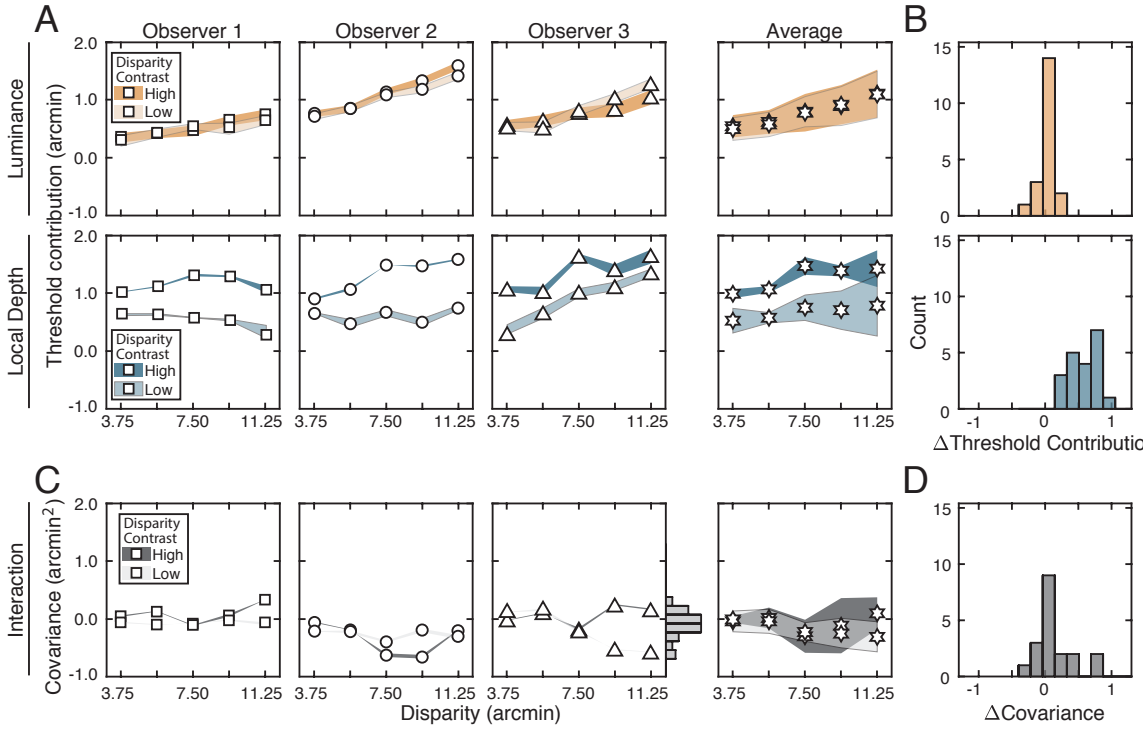

Figure 10. Contributions of distinct stimulus-specific factors to thresholds, as revealed by the quasi-quadruple-pass analysis. **A**. Contribution of luminance-contrast pattern variability (*top*) and variability in local-depth structure (*bottom*) to threshold as a function of disparity pedestal at different disparity-contrast levels (*shades*), for each observer and the observer average. For individual observers, bounds of shaded regions indicate 68% confidence intervals for each condition, generated from 10,000 bootstrapped samples. For the observer average, bounds indicate across-observer standard-deviations. **B**. Histogram of differences in luminance-pattern-driven and local-depth-driven threshold contributions across high and low disparity-contrast conditions, collapsed across disparity pedestals and individual observers. **C**. Same as *A*, but for the interaction term (i.e. $\text{cov}[L, B]$). Histogram of the interaction terms collapsed across all disparity pedestals, disparity-contrasts, and individual observers is shown on the rightmost y-axis of the third column (mean=-0.11, sd=0.23). **D**. Histogram of differences in the interaction term (i.e. $\text{cov}[L, B]$) across high and low disparity-contrast conditions, collapsed across disparity pedestals and individual observers. Data in *C*-*D* indicate that the interaction term is near-zero in all conditions.

stimuli, stimulus-by-stimulus estimates or trial-by-trial responses from one observer will provide no information about estimates or responses from another, and between-subjects decision-variable correlation should equal zero.

We computed between-observers decision-variable correlation from response agreement data by straightforward adaptation of the quasi-quadruple-pass analysis (see Methods). However, because between-observers correlation is impacted by internal noise, its value does not transparently reflect the level of shared stimulus drive. The partial correlation does. Partial correlation is given by

$$\rho_{12 \cdot W} = \frac{\rho_{12}}{\sqrt{\rho_{11}\rho_{22}}} = \frac{\mathrm{cov}[S_1, S_2]}{\sigma_{E1}\sigma_{E2}}, \tag{51}$$

where ρ_{12} is between-observers decision-variable correlation, ρ_{11} and ρ_{22} are the within-observer decision-variable correlations, S_1 and S_2 are the stimulus-driven components of the decision variable that are shared between the two observers, and σ_{E1} and σ_{E2} are the standard-deviations of the stimulus-driven components of the decision variable in the two observers. This partial correlation provides more unvarnished information about what we are most interested in, because it is unaffected by internal noise. It quantifies the level of correlation in the stimulus-driven component of the decision variable between observers (see Methods).

Between-observers partial correlations are shown in Figure 11. Across all conditions and observer pairs, between-observers partial correlations are substantially above zero. In the high disparity-contrast conditions of Experiment 1, which are the conditions in which local-depth variability has its largest effects, between-observers partial correlations are 0.79 on average, with some values approaching the maximum possible value (i.e. 1.0). In the low disparity-contrast conditions of Experiment 1, the average value is 0.59. In Experiment 2, the average partial correlations for the high and low disparity-contrast conditions are 0.56 and 0.53, respectively (Fig. 11 bottom row). Histograms of the differences between the high- and low-disparity-contrast conditions are shown in Figure 11B. And histograms of the raw values are shown in Figure 11C.

These results indicate that the majority—and, in one case, the strong majority—of the stimulus-driven component of the decision variable is shared between observers. That is, natural stimulus variability associated with different stimuli having same the latent variable (i.e. disparity) causes similar stimulus-by-stimulus over- and under-estimations of disparity-defined-depth in different humans. We conclude that the deterministic computations that the human visual system performs on individual stimuli are largely consistent across observers.

Chin and Burge (2020), in the the domain of speed discrimination, came to a similar conclusion using

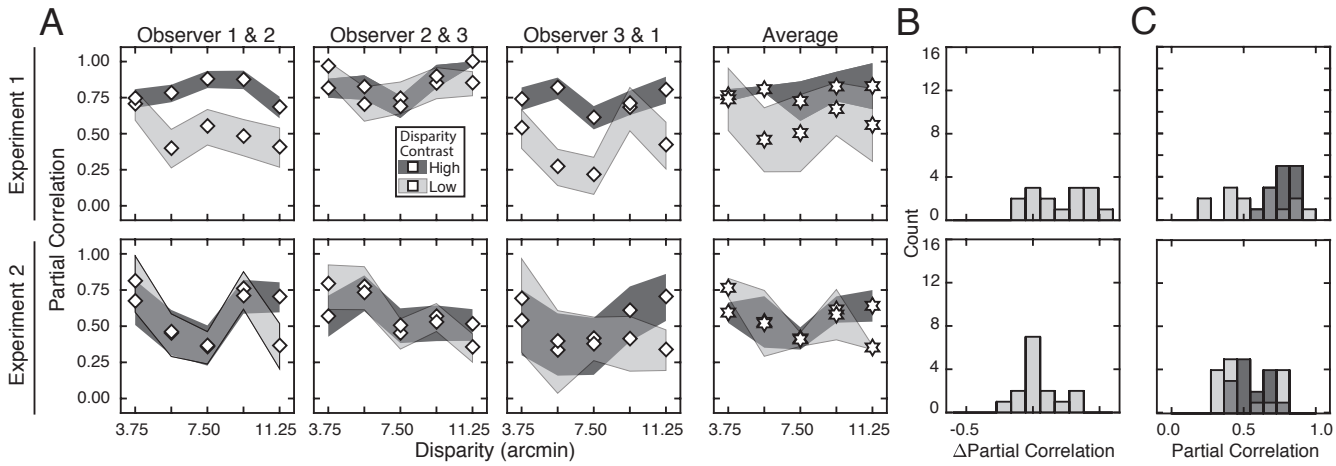

Figure 11. Between-observer correlation in the stimulus-driven component of the decision variable, as revealed by the quasi-quadruple-pass analysis. **A**. Estimated partial correlation values, controlling for (i.e. removing) the affect of internal noise, between all observer pairs, for each experiment (*rows*), at all disparity and disparity-contrast levels. Averages across observer-pairs are shown in *column 4*. With the effect of internal noise removed, only the stimulus-driven component of the decision variable drives between-observer correlation. For observer pairs, bounds of shaded regions indicate 95% confidence intervals for each condition from 1,000 bootstrapped datasets. For the across-observer-pair average, bounds of shaded region indicates across-pair standard deviations. **B**. Histogram of differences in partial correlation across high- and low-disparity-contrast conditions shown in *A*, collapsed across disparity pedestals and observer pairs. **C**. Histograms of the raw partial correlations for each observer pair in *A*. In Experiment 1, the mean partial correlations are 0.79 and 0.59 in the high- and low-disparity-contrast conditions, respectively. In Experiment 2, the values are 0.56 and 0.53. The majority of the stimulus-driven variance is shared between observers.

a related approach. By comparing human performance to that of an image-computable ideal observer, they found that differing levels of human inefficiency are near-exclusively attributable to different levels of internal noise. Like the current findings, this finding entailed that the variance of the stimulus-driven component of the decision variable is quite similar across different human observers, and is consistent with the visual systems of different human observers performing the same deterministic computations on the stimuli. The dovetailing evidence from the current study of disparity-based depth discrimination and the previous study of speed discrimination suggest that natural stimulus variability (natural variation in luminance pattern and/or depth-structure) has consistent effects on the visual systems of different human observers. These results suggest that evolution has tightly honed the details of how visual systems compute so that they extract the most useful task-relevant information from natural stimuli.

6 Discussion

In this article, using a natural-stimulus dataset, two double-pass experiments, and a series of analyses, we investigated human stereo-depth discrimination in natural scenes, with specific emphasis on how natural-stimulus variability limits performance. We sourced stimuli from a natural stereo-image database with a constrained sampling procedure, and computed ground-truth disparities directly from laser-range data at each pixel. Fixation (or pedestal) disparity, and local-depth variability—as quantified by disparity-contrast—were tightly controlled. Luminance-contrast patterns and local-depth structures were allowed to vary naturally across the hundreds of unique stimuli that were sampled for each condition.

We find that the exponential law of disparity discrimination holds for human vision in natural scenes. We find that stimulus-driven variability and noise-driven variability have near-equal roles in setting these thresholds, and that the stimulus-based sources of variability make responses more repeatable (and thus potentially more predictable) across repeated stimulus-presentations. We find that one of two underlying causes of the stimulus-driven variability is attributable to local-depth variation, multiplicatively increases discrimination thresholds, and is largely separable from luminance-contrast-pattern variation. And we find that different subjects make correlated stimulus-by-stimulus over- and under-estimations of disparity, suggesting that the different human visual systems process individual natural stimuli with computations that are largely the same.

The approach developed here extends the rigor and interpretability that has been integral to progress in more traditional psychophysics and neuroscience experiments to more natural-stimulus sets Sebastian et al., 2017; Chin and Burge, 2020; Ziemba et al., 2016. In the real world, perceptual, and behavioral

variability is driven by both external and internal factors. A comprehensive account of perceptual and behavioral variability, and the neural activity underlying it, must identify and describe the impact of all significant sources of performance-limiting variability. Encouragingly, the current results raise the prospect that an appropriate image-computable model may, in principle, be able to predict a substantial proportion of stimulus-by-stimulus variation across natural images.

6.1 Progress and limitations

Progress in science is often incremental. Many times, it occurs by way of relaxing one experimental design element, while holding others fixed. We and others have investigated perceptual performance with stimuli sampled from natural scenes—which are atypical of laboratory experiments—, while using conventional, tightly controlled, laboratory tasks (Burge & Geisler, 2015; Sebastian et al., 2017; Chin & Burge, 2020; Ziemba et al., 2016). Others have investigated performance with atypical tasks (e.g. free viewing and unconstrained eye-movements), while using conventional (e.g. Gabor) stimuli (Yates et al., 2023). Both approaches have increased the ecological validity of the experimental conditions, and have provided new insights into the properties of neural computations underlying sensory-perceptual performance. But there are always limitations.

The stimuli used in the current experiments were foveally presented and subtended only 1° of visual angle, the approximate size of foveal receptive fields in early visual cortex. Foveal presentation of spatially-limited stimuli is common in psychophysical experiments, but doing so prevents the assessment of peripheral visual processing or how performance is affected by the dynamic interplay between eye, head, and body movements occurring in natural viewing. Limiting stimulus size to one degree also limits the degree to which contextual effects can affect performance.

Related issues concern the two-alternative forced choice (2AFC) procedure used in the current experiments. Although commonly employed, the rigid trial structure imposed by such designs is not well-aligned with how perceptual estimates, perception-driven decisions, and perception-guided action are inter-related in natural viewing. Alternative methods, such as continuous psychophysics, that more closely reflect the continuous interplay of perception and action in natural viewing, could complement the current findings (Bonnen et al., 2015; Chin & Burge, 2022; Burge & Cormack, 2020).

Despite these limitations, the current experiments showed that the natural variation of luminance-contrast patterns and local-depth structures have large, distinct, and identifiable effects on performance. Developing methods that guide the judicious choice of stimulus sets and tasks that strike an appropriate

balance between fully natural and tightly constrained, that are well-suited to available analytical methods, and are well-matched to the specific research question under study, will be increasingly important as the science becomes more focused on understanding how neurons respond and how perception works in the natural environment.

6.2 Performance variation and prediction

An ultimate goal of perception science is to be able to predict, from an individual stimulus, the neural activity and subsequent perceptual estimate, whether it will be accurate or inaccurate, and whether it will be reliable or unreliable. The degree to which this goal is achievable hinges on the degree to which the stimulus-by-stimulus estimates are controlled by the properties of the stimulus, as opposed to noise. If the strong majority of performance variation is noise-driven, such efforts will be futile. So, before undertaking to develop such models, it is prudent to demonstrate that a substantial proportion of performance variation is stimulus driven. In the current stereo-depth discrimination experiments with natural stimuli, stimulus-based sources of response variability account for approximately half of all performance-limiting variability in the results reported here (see Fig. 4), and a substantial proportion of that stimulus-driven variability was shared across observers (see Fig. 11).

However, while the stimuli were allowed to vary naturally in many respects—they were, of course, stereo-photographs of natural scenes—the mean luminance was fixed to a comfortable photopic level, and luminance-contrast was set to the median contrast in natural scenes (see Methods) (A. Iyer & Burge, 2019; Frazor & Geisler, 2006). Both properties are known to impact stereo-depth discrimination performance (Cormack et al., 1991), and detection performance more generally (Sebastian et al., 2017; Mueller, 1951; Nachmias & Sansbury, 1974; Legge & Foley, 1980; Burgess et al., 1981). Indeed, as mean-luminance and luminance-contrast increase, neurons respond more vigorously, that signal-to-noise ratios increase, and that performance becomes more reliable (Frazor & Geisler, 2006; Mante et al., 2005). Hence, if luminance and contrast had been allowed to vary more naturally, the proportional contribution of stimulus-based factors to performance-limiting variability may increase. The current estimates of stimulus-based contributions to the decision variable may therefore be underestimates of the total impact that stimulus-based factors would have in less tightly controlled circumstances. The fact that the average between-observers partial correlation are near the maxima in the conditions in which natural stimulus variability was highest, lends credence to the speculation (see Fig. 11).

In light of these considerations, experimental designs should present unique stimuli on each trial, because

doing so will increase the power of empirical datasets to help develop, constrain, and evaluate models. Many models can yield similar predictions of performance if only summary statistics (e.g. bias and precision) are used to evaluate the models' successes and failures. Image-computable models that predict decision-variable correlation and stimulus-by-stimulus estimates (or discriminations), in addition to bias and/or precision, can provide increased power for evaluating hypotheses about the neural activity and sensory-perceptual computations underlying performance (Burge, 2020; Chin & Burge, 2020; Geisler, 2018).

6.3 Noise and its impact on performance

In this article, we sought to partition the influence on performance of stimulus-driven from noise-driven variability, and to further partition the effects of two distinct types of natural-stimulus variability: luminance-pattern and local-depth variability. We made no attempts to determine different potential sources of noise (i.e. stimulus-independent sources variability), or to partition the influence of each on performance. As a consequence, any source of variance that led to less repeatable responses in the current experiments contributed to the estimate of noise variance. We conceptualized the noise as occurring at the level of the decision variable. But there are multiple stages in the chain of events preceding perceptual estimatation, both external and internal to the organism, where such variability could have originated and that would be consistent with the results.

Variation due to noise could have occurred during the initial encoding of the retinal image, in early visual cortex, at the decision stage (e.g. in the placement of the criterion), or a combination of these possibilities. Potential sources of such variation include the noisy nature of light itself (Hecht, 1942), random fixational errors (Ukwade et al., 2003), neural noise (Tomko & Crapper, 1974; Tolhurst et al., 1983), and trial-sequential dependencies (Laming, 1979). Higher-level factors could also manifest as noise, including stimulus-independent fluctuations in alertness, attention, or motivation (Lu & Dosher, 1998; Mitchell et al., 2009; J. I. Gold & Ding, 2013; Zhang et al., 2018).

Experimental and computational methods that can determine the contribution of different types of stimulus-independent sources of variation are of interest to systems neuroscience (Chin & Burge, 2020; Goris et al., 2014). There are clear steps that could be taken to identify and account for some of these potential sources of noise. Psychophysical methods have the potential to distinguish some of them. High-resolution eye-tracking would allow one to condition performance on the fixational state of the eyes (Bowers et al., 2019; S. Stevenson et al., 2016; Rucci & Poletti, 2015). Parametrically varying performance-contingent reward can systematically alter motivational state (Zhang et al., 2018). But neurophysiological methods

would be required to identify and partition sources of noise internal to the nervous system that may arise at various stages of the visual processing and perceptual decision making pipeline. Paradigms that blend the advantages of the current approach for partitioning stimulus-based variation with neurophysiological and computational methods for partitioning noise would be a useful way forward (Ziemba et al., 2016; Goris et al., 2014; Charles & Pillow, 2018).

6.4 External limits to human performance

Broadly construed, the current work continues in the tradition of the classic 1942 study of Hecht, Shlaer, and Pirenne. Its two most widely appreciated results are that, when fully dark adapted, i) the absorption of a single photon reliably elicits a response from a rod photoreceptor and ii) the absorption of five to seven photons in a short period of time reliably causes a reportable sensation of the light. Slightly less widely appreciated is the finding that the limits of the human ability to detect light (i.e. light detectability thresholds) are attributable to the stochastic nature of light itself, a performance-limiting factor that is external to the organism. On a given trial at a given stimulus intensity, whether or not subjects reported that they had seen the stimulus depended near-exclusively on whether or not the requisite number of photons had been absorbed. That is, if the numbers of photons in proximal stimulus was identical, humans would respond identically. Performance was thus very tightly yoked to the variability of the external stimulus.

The results the current study suggest that, just as rod photoreceptors support performance in a very similar manner across different human observers, the computational mechanisms supporting the estimation and discrimination of depth in natural scenes are very similar across observers. In the current study, we showed that stimulus-based limits to performance become increasingly important as stimuli become ever more natural. If this pattern holds, it may that stimulus-based limits to performance are by far the dominant factor as organisms engage with the natural environment. If true, image-computable models will have the potential to achieve remarkable predictive power from analysis of the stimulus alone. Such models, in which the underlying computations are made explicit, would have tremendous practical applications and deepen our understanding of how vision works in the real world.

References

Baddeley, R. (1997). The Correlational Structure of Natural Images and the Calibration of Spatial Representations. *Cognitive Science*, *21*(3), 351–372. https://doi.org/10.1207/s15516709cog2103_4

Felsen, G., & Dan, Y. (2005). A natural approach to studying vision. *Nature Neuroscience*, *8*(12), 1643–1646. https://doi.org/10.1038/nn1608

Iyer, A., & Burge, J. (2019). The statistics of how natural images drive the responses of neurons. *Journal of Vision*, *19*(13), 4. https://doi.org/10.1167/19.13.4

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, *96*(2), 267–314. https://doi.org/10.1037/0033-295X.96.2.267

Burge, J., & Geisler, W. S. (2011). Optimal defocus estimation in individual natural images. *Proceedings of the National Academy of Sciences*, *108*(40), 16849–16854. https://doi.org/10.1073/pnas.1108491108

Burge, J., & Geisler, W. S. (2012, January 22). Optimal defocus estimates from individual images for autofocusing a digital camera. In S. Battiato, B. G. Rodricks, N. Sampat, F. H. Imai, & F. Xiao (Eds.). https://doi.org/10.1117/12.912066

Burge, J., & Geisler, W. S. (2014). Optimal disparity estimation in natural stereo images. *Journal of Vision*, *14*(2), 1–1. https://doi.org/10.1167/14.2.1

Burge, J., & Geisler, W. S. (2015). Optimal speed estimation in natural image movies predicts human performance. *Nature Communications*, *6*(1), 7900. https://doi.org/10.1038/ncomms8900

Sebastian, S., Burge, J., & Geisler, W. S. (2015). Defocus blur discrimination in natural images with natural optics. *Journal of Vision*, *15*(5). https://doi.org/10.1167/15.5.16

Sebastian, S., Abrams, J., & Geisler, W. S. (2017). Constrained sampling experiments reveal principles of detection in natural scenes. *Proceedings of the National Academy of Sciences*, *114*(28), E5731–E5740. https://doi.org/10.1073/pnas.1619487114

Burge, J. (2020). Image-Computable Ideal Observers for Tasks with Natural Stimuli. *Annual Review of Vision Science*, *6*(1), 491–517. https://doi.org/10.1146/annurev-vision-030320-041134

Chin, B. M., & Burge, J. (2020). Predicting the Partition of Behavioral Variability in Speed Perception with Naturalistic Stimuli. *The Journal of Neuroscience*, *40*(4), 864–879. https://doi.org/10.1523/JNEUROSCI.1904-19.2019

Burgess, A. E., & Colborne, B. (1988). Visual signal detection IV Observer inconsistency. *Journal of the Optical Society of America A*, *5*(4), 617. https://doi.org/10.1364/JOSAA.5.000617

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Signal but not noise changes with perceptual learning. *Nature*, *402*(6758), 176–178. https://doi.org/10.1038/46027

Neri, P., & Levi, D. M. (2006). Receptive versus perceptive fields from the reverse-correlation viewpoint. *Vision Research*, *46*(16), 2465–2474. https://doi.org/10.1016/j.visres.2006.02.002

Burge, J., McCann, B. C., & Geisler, W. S. (2016). Estimating 3D tilt from local image cues in natural scenes. *Journal of Vision*, *16*(13), 2. https://doi.org/10.1167/16.13.2

Iyer, A. V., & Burge, J. (2018). Depth variation and stereo processing tasks in natural scenes. *Journal of Vision*, *18*(6), 4–4. https://doi.org/10.1167/18.6.4

Held, R. T., & Banks, M. S. (n.d.). Misperceptions in Stereoscopic Displays: A Vision Science Perspective, 10.

Blakemore, C. (1970). The range and scope of binocular depth discrimination in man. *The Journal of Physiology*, *211*(3), 599–622. https://doi.org/10.1113/jphysiol.1970.sp009296

Cormack, L. K., Stevenson, S. B., & Schor, C. M. (1991). Interocular correlation, luminance contrast and cyclopean processing. *Vision Research*, *31*(12), 2195–2207. https://doi.org/10.1016/0042-6989(91)90172-2

Badcock, D. R., & Schor, C. M. (1985). Depth-increment detection function for individual spatial channels. *Journal of the Optical Society of America A*, *2*(7), 1211. https://doi.org/10.1364/JOSAA.2.001211

McKee, S. P., Levi, D. M., & Bowne, S. F. (1990). The imprecision of stereopsis. *Vision Research*, *30*(11), 1763–1779. https://doi.org/10.1016/0042-6989(90)90158-H

Stevenson, S. B., Cormack, L. K., Schor, C. M., & Tyler, C. W. (1992). Disparity tuning in mechanisms of human stereopsis. *Vision Research*, *32*(9), 1685–1694. https://doi.org/10.1016/0042-6989(92)90161-B

Banks, M. S. (2004). Why Is Spatial Stereoresolution So Low? *Journal of Neuroscience*, *24*(9), 2077–2089. https://doi.org/10.1523/JNEUROSCI.3852-02.2004

Nienborg, H., Bridge, H., Parker, A. J., & Cumming, B. G. (2004). Receptive Field Size in V1 Neurons Limits Acuity for Perceiving Disparity Modulation. *The Journal of Neuroscience*, *24*(9), 2065–2076. https://doi.org/10.1523/JNEUROSCI.3887-03.2004

Tyler, C. W. (1974). Depth perception in disparity gratings. *Nature*, *251*(5471), 140–142. https://doi.org/10.1038/251140a0

Westheimer, G. (1979). Cooperative neural processes involved in stereoscopic acuity. *Experimental Brain Research*, *36*(3). https://doi.org/10.1007/BF00238525

Ziemba, C. M., Freeman, J., Movshon, J. A., & Simoncelli, E. P. (2016). Selectivity and tolerance for visual texture in macaque V2. *Proceedings of the National Academy of Sciences*, *113*(22), E3140–E3149. https://doi.org/10.1073/pnas.1510847113

Yates, J. L., Coop, S. H., Sarch, G. H., Wu, R.-J., Butts, D. A., Rucci, M., & Mitchell, J. F. (2023). Detailed characterization of neural selectivity in free viewing primates. *Nature Communications*, *14*(1), 3656. https://doi.org/10.1038/s41467-023-38564-9

Bonnen, K., Burge, J., Yates, J., Pillow, J., & Cormack, L. K. (2015). Continuous psychophysics: Target-tracking to measure visual sensitivity. *Journal of Vision*, *15*(3), 14. https://doi.org/10.1167/15.3.14

Chin, B. M., & Burge, J. (2022). Perceptual consequences of interocular differences in the duration of temporal integration. *Journal of Vision*, *22*(12), 12. https://doi.org/10.1167/jov.22.12.12

Burge, J., & Cormack, L. K. (2020, August 6). *Target tracking reveals the time course of visual processing with millisecond-scale precision* (preprint). Neuroscience. https://doi.org/10.1101/2020.08.05.238642

Frazor, R. A., & Geisler, W. S. (2006). Local luminance and contrast in natural images. *Vision Research*, *46*(10), 1585–1598. https://doi.org/10.1016/j.visres.2005.06.038

Mueller, C. G. (1951). Frequency of seeing functions for intensity discrimination at various levels of adapting intensity. *Journal of General Physiology*, *34*(4), 463–474. https://doi.org/10.1085/jgp.34.4.463

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

Legge, G. E., & Foley, J. M. (1980). Contrast masking in human vision. *Journal of the Optical Society of America*, *70*(12), 1458. https://doi.org/10.1364/JOSA.70.001458

Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. (1981). Efficiency of Human Visual Signal Discrimination. *Science*, *214*(4516), 93–94. https://doi.org/10.1126/science.7280685

Mante, V., Frazor, R. A., Bonin, V., Geisler, W. S., & Carandini, M. (2005). Independence of luminance and contrast in natural scenes and in the early visual system. *Nature Neuroscience*, *8*(12), 1690–1697. https://doi.org/10.1038/nn1556

Geisler, W. S. (2018). Psychometric functions of uncertain template matching observers. *Journal of Vision*, *18*(2), 1. https://doi.org/10.1167/18.2.1

Hecht, S. (1942). Energy, quanta, and vision. *The Journal of General Physiology*, *25*(6), 819–840. https://doi.org/10.1085/jgp.25.6.819

Ukwade, M. T., Bedell, H. E., & Harwerth, R. S. (2003). Stereothresholds with simulated vergence variability and constant error. *Vision Research*, *43*(2), 195–204. https://doi.org/10.1016/S0042-6989(02)00409-1

Tomko, G. J., & Crapper, D. R. (1974). Neuronal variability: Non-stationary responses to identical visual stimuli. *Brain Research*, *79*(3), 405–418. https://doi.org/10.1016/0006-8993(74)90438-7

Tolhurst, D., Movshon, J., & Dean, A. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research*, *23*(8), 775–785. https://doi.org/10.1016/0042-6989(83)90200-6

Laming, D. (1979). Choice reaction performance following an error. *Acta Psychologica*, *43*(3), 199–224. https://doi.org/10.1016/0001-6918(79)90026-X

Lu, Z.-L., & Dosher, B. A. (1998). External noise distinguishes attention mechanisms. *Vision Research*, *38*(9), 1183–1198. https://doi.org/10.1016/S0042-6989(97)00273-3

Mitchell, J. F., Sundberg, K. A., & Reynolds, J. H. (2009). Spatial Attention Decorrelates Intrinsic Activity Fluctuations in Macaque Area V4. *Neuron*, *63*(6), 879–888. https://doi.org/10.1016/j.neuron.2009.09.013

Gold, J. I., & Ding, L. (2013). How mechanisms of perceptual decision-making affect the psychometric function. *Progress in Neurobiology*, *103*, 98–114. https://doi.org/10.1016/j.pneurobio.2012.05.008

Zhang, P., Hou, F., Yan, F.-F., Xi, J., Lin, B.-R., Zhao, J., Yang, J., Chen, G., Zhang, M.-Y., He, Q., Dosher, B. A., Lu, Z.-L., & Huang, C.-B. (2018). High reward enhances perceptual learning. *Journal of Vision*, *18*(8), 11. https://doi.org/10.1167/18.8.11

Goris, R. L. T., Movshon, J. A., & Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience*, *17*(6), 858–865. https://doi.org/10.1038/nn.3711

Bowers, N. R., Gibaldi, A., Alexander, E., Banks, M. S., & Roorda, A. (2019). High-resolution eye tracking using scanning laser ophthalmoscopy. *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 1–3. https://doi.org/10.1145/3314111.3322877

Stevenson, S., Sheehy, C., & Roorda, A. (2016). Binocular eye tracking with the Tracking Scanning Laser Ophthalmoscope. *Vision Research*, *118*, 98–104. https://doi.org/10.1016/j.visres.2015.01.019

Rucci, M., & Poletti, M. (2015). Control and Functions of Fixational Eye Movements. *Annual Review of Vision Science*, *1*(1), 499–518. https://doi.org/10.1146/annurev-vision-082114-035742

Charles, A., & Pillow, J. (2018). Additive Continuous-time Joint Partitioning of Neural Variability. *2018 Conference on Cognitive Computational Neuroscience*. https://doi.org/10.32470/CCN.2018.1087-0