Probabilistic combination of slant information: Weighted averaging and robustness as optimal percepts

Ahna R. Girshick
Department of Psychology and Center for Neural Science,
New York University, New York, NY, USA

Martin S. Banks
Vision Science Program, Department of Psychology and
Helen Wills Neuroscience Institute,
University of California, Berkeley, CA, USA

Depth perception involves combining multiple, possibly conflicting, sensory measurements to estimate the 3D structure of the viewed scene. Previous work has shown that the perceptual system combines measurements using a statistically optimal weighted average. However, the system should only combine measurements when they come from the same source. We asked whether the brain avoids combining measurements when they differ from one another: that is, whether the system is robust to outliers. To do this, we investigated how two slant cues—binocular disparity and texture gradients—influence perceived slant as a function of the size of the conflict between the cues. When the conflict was small, we observed weighted averaging. When the conflict was large, we observed robust behavior: perceived slant was dictated solely by one cue, the other being rejected. Interestingly, the rejected cue was either disparity or texture, and was not necessarily the more variable cue. We modeled the data in a probabilistic framework, and showed that weighted averaging and robustness are predicted if the underlying likelihoods have heavier tails than Gaussians. We also asked whether observers had conscious access to the single-cue estimates when they exhibited robustness and found they did not, i.e. they completely fused despite the robust percepts.

Keywords: cue combination, robust estimation, visual perception, slant estimation, heavy-tailed estimators, sensory fusion

Citation: Girshick, A. R., & Banks, M. S. (2009). Probabilistic combination of slant information: Weighted averaging and robustness as optimal percepts. *Journal of Vision, 9*(9):8, 1–20, http://journalofvision.org/9/9/8/, doi:10.1167/9.9.8.

Introduction

In 1882, Simon Newcomb, a Canadian-American astronomer and mathematician, developed a refinement of Foucault's method for measuring the speed of light. With his new method, he made 66 repeated measurements of the time required for light to travel 7442 m. The data set had a large cluster of measurements that were approximately Gaussian distributed, but there were also two unusually low measurements. The mean of the 66 measurements was 2.4826×10^{-5} sec, which corresponds to a speed of 2.9976×108 m/sec. However, if the two low values were first rejected, the mean was 2.4828×10^{-5} sec, which is closer to the currently accepted value of 2.4833×10^{-5} for Newcomb's experiment (Gelman, Carlin, Stern, & Rubin, 2003). In the analysis of data sets like Newcomb's, the mean is often used to estimate the location of the center of the data, but outliers can have a large and potentially detrimental effect by dragging the mean away from the bulk of the data. It can be useful, therefore, to adjust the data by eliminating or down-weighting outliers before computing the statistic of interest. Robust statistics provide methods to do just that; the methods allow the estimation of statistics such as the central tendency and variation without being unduly affected by outliers (Huber, 1981). In this paper, we examine the visual system's treatment of sensory information that is either relatively consistent or quite inconsistent. We examine in particular whether the system exhibits behavior similar to statistical robustness.

The human perceptual system uses various sources of sensory information, prior expectations, and expected rewards and costs in a fashion that is often consistent with Bayesian inference (Knill, Kersten, & Yuille, 1996). In most cases, sensory measurements seem to be combined in a weighted average with the weights proportional to the normalized reliability of each measurement (Ernst & Banks, 2002). But weighted averaging does not necessarily occur when the measurements are quite discrepant from one another (van Ee, van Dam, & Erkelens, 2002). Here we investigate measurement combination when the discrepancy is small and large, and we ask whether a general rule can be established. Understanding this issue is relevant to understanding sensory combination for a wide variety of cases including within-modality signals (e.g., depth from stereo and from motion), between-modality signals (e.g., position from vision and audition), and the influence of *a priori* information (e.g., that light generally comes from above).

doi: 10.1167/9.9.8 Received April 19, 2009; published August 24, 2009 ISSN 1534-7362 © ARVO

Bayes' rule, weighted averaging, and cue combination with Gaussian likelihoods

The Bayesian model with Gaussian likelihoods has become the standard framework for sensory cue combination (Ghahramani, Wolpert, & Jordan, 1997; Landy, Maloney, Johnston, & Young, 1995; Yuille & Bülthoff, 1996), and is supported by a large body of empirical evidence (Alais & Burr, 2004; Ernst & Banks, 2002; Hillis, Watt, Landy, & Banks, 2004; Jacobs, 1999; Knill & Saunders, 2003). Let S denote an environmental variable (e.g., shape, size, or slant) and let $\{X_i\}$ for i = 1, …, N denote the perceptual system's measurements from N cues. All sensory measurements are subject to variation due to measurement error and variation in the mapping between the environment and sensory apparatus. It is conventionally assumed that X_i are Gaussian distributed with variance σ_i^2 and are conditionally independent (i.e., their noises are independent). Using Bayes' rule and the assumption that X_i are conditionally independent, the brain can minimize the uncertainty about the environmental variable:

$$p(S|X_1, \ldots X_N) \propto p(X_1|S) \ldots p(X_N|S)p(S), \tag{1}$$

where X_i is the sensor image data corresponding to the ith cue. The first N terms on the right side of the equation are the likelihood functions representing the probabilities of observing the sensor data from each of the N measurements if S is the actual value of the environmental variable. The last term is the prior distribution, the probability of observing the value S in the scene; it is independent of the sensory data. The left side is the posterior distribution representing the combined estimate from the measurements. Unless there are immediate consequences to certain actions (payoffs and penalties), it is most advantageous for the observer to choose the value of S that maximizes the posterior: this is the maximum *a posteriori* (MAP) rule.

In many cases, the prior is broad and has a negligible effect on the combined estimate. In the specific case considered here, the standard deviation of the prior distribution for slant (~40°) is likely to be much larger than the standard deviation of the likelihoods associated with disparity and texture (Arnold & Binford, 1980; Hillis et al., 2004), so we will ignore it in the remainder of our analysis. We will use the phrase *Gaussian-likelihood model* to refer to the case in which the likelihoods are Gaussian and conditionally independent. Assuming a sensible decision rule, such as MAP, this model always predicts *weighted averaging* (a phrase that derives from expressing the likelihoods in Equation 1 as Gaussian and their product as a weighted sum of Gaussian-distributed random variables). In weighted averaging, the combined estimate is always in-between the single-cue estimates, which is represented schematically in Figure 1b.

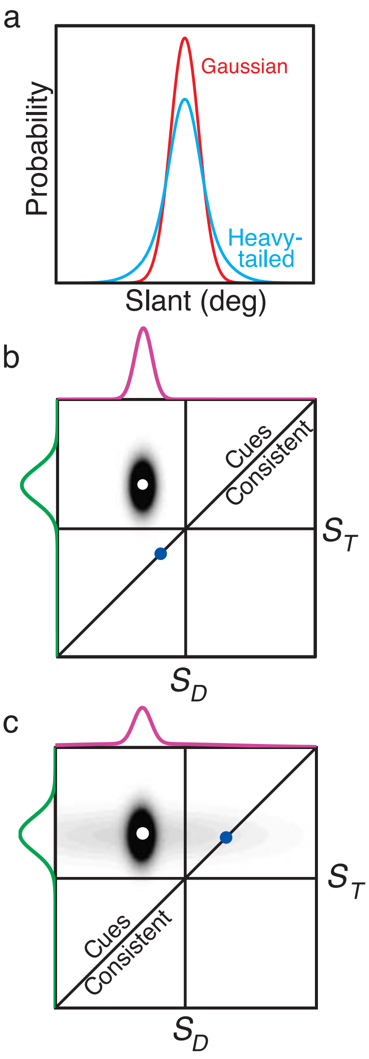

Figure 1. a) Gaussian (red) and heavy-tailed (blue) likelihood functions. b–c) The joint likelihoods with disparity-specified slant, S_D, on the horizontal axis and texture-specified slant, S_T, on the vertical axis, assumed to be conditionally independent. Points on the cues-consistent (diagonal) line represent stimuli with the same disparity- and texture-specified slants, whereas points off this line represent stimuli with conflicting slants. Cue-conflict size increases as points get farther from the cues-consistent line. Profiles of the likelihood functions are shown for disparity (above, pink, smaller variance) and texture (left, green, larger variance). Their joint likelihood functions are depicted by gray clouds for which probability (calculated using Equation 1) is proportional to intensity; white dots mark the peaks. The blue dots are the *cues-consistent predictions*, calculated as the peaks of the intersection profiles of the joint likelihood functions and the cues-consistent axis. b) Gaussian likelihood functions create a joint likelihood function whose profile is elliptical. The cues-consistent prediction is the weighted average as determined by the relative reliabilities of the two cues. c) Heavy-tailed likelihood functions create a joint likelihood function whose profile forms a cross. The cues-consistent prediction is robust and chooses texture because, even though the texture likelihood is more variable, it has less heavy tails.

Large cue conflicts, heavy-tailed likelihood functions, and cue combination

If sensory measurements are quite discrepant, combining them may be misguided; it depends on the cause of the discrepancy. There are cases in which combining would yield an erroneous estimate: if the measurements come from different objects or parts of an object, if one or more of the measurements comes from a faulty sensor, or if the wrong generative model has been used to interpret a measurement (e.g., assuming a trapezoidal texture instead of a square one). But there are also cases in which combining would yield a more accurate estimate: i.e., cases in which the discrepancy is due to unbiased random error that affects the measurements. The magnitude of the conflict between measurements is likely to be correlated with the cause of the conflict: larger conflicts being more likely to have been caused by measurement from different objects, a faulty sensor, or an incorrect generative model. It makes sense then that the size of the conflict would be a determining factor in whether to combine or not (Ernst, 2005; Körding et al., 2007; Natarajan, Murray, Shams, & Zemel, 2009; Sato, Toyoizumi, & Aihara, 2007). As in our initial example of outlying measurements, the visual system can treat a conflicting signal as an outlier. This leads to our central question: when faced with quite conflicting signals, does the visual system continue to do weighted averaging, or does it down-weight a cue and thereby behave as a robust estimator?

With N measurements (1 from each of N cues), one outlier may be obvious among $N - 1$ concurring measurements; standard techniques in robust statistics proscribe how much to down-weight that measurement (Gelman et al., 2003; Huber, 1981). Here we investigate the situation in which $N = 2$, a case in which standard statistical techniques do not pinpoint the outlier.

Situations in which a cue is ignored in favor of another are well documented in perception. Such situations are referred to as 'cue dominance' (Howard & Templeton, 1966). Cue dominance is exemplified by 'visual capture' in visual-haptic perception (Hay & Pick, 1966; Rock & Victor, 1964) and in visual-auditory perception (Pick, Warren, & Hay, 1969). An important question is whether cue dominance occurs simply because one cue (i.e., vision) always dominates when two are in conflict, or whether such behavior emerges from a more general and statistically sensible process. This issue has recently been examined in a probabilistic framework, which revealed that apparent visual dominance over haptics and audition simply occurs when vision is statistically more reliable (Alais & Burr, 2004; Ernst & Banks, 2002; Gepshtein & Banks, 2003). In these cases, the dominance observations are attributable to weighted averaging of cues with weights near 0 and 1.

There are also numerous demonstrations in which two cues affect the percept when the cue conflict is small, but one cue is ignored when the conflict reaches a critical value; this has been called 'cue vetoing' (Bülthoff & Mallot, 1988). An example is the induced effect (Ogle, 1938). In this effect, a vertical magnifier is placed before one eye, which alters the vertical but not the horizontal disparities associated with a viewed surface. With the magnifier in place, a fronto-parallel plane appears rotated about a vertical axis. As the magnification is increased from 0 to 5%, the plane's apparent slant increases monotonically. With further magnification increases, however, apparent slant regresses to zero. The perception of non-zero slant with small magnifications is consistent with weighted averaging of the available slant signals; the regression to zero slant with large magnifications is consistent with the vetoing of one slant signal (vertical disparity coupled with horizontal disparity; Banks & Backus, 1998). There are other phenomena in which the percept follows one cue or another, and apparently never adopts an average including the Necker Cube (Necker, 1832) and reverse-perspective paintings (Wade & Hughes, 1999).

How might robustness occur in the probabilistic framework? Recall that cue combination is implemented in this framework as the product of the likelihood functions. Previous research has assumed that the likelihoods are Gaussian with means μ_i and variances σ_i^2. The product of two Gaussians is a Gaussian with mean $(\mu_1\sigma_2^2 + \mu_2\sigma_1^2)/(\sigma_1^2 + \sigma_2^2)$ and variance $\sigma_1^2\sigma_2^2/(\sigma_1^2 + \sigma_2^2)$. As the means of two Gaussians become increasingly disparate, the product's mean always remains at a fixed proportional distance in-between the two, corresponding to the weighted-average prediction. Thus, in this framework, Gaussian likelihood functions can never produce robust behavior, even at very large cue conflicts. However, small changes in the shape of the likelihood functions can yield significant changes in the position and shape of the posterior for large conflicts and thereby produce robust behavior (Box & Tiao, 1992; Knill, 2003). If the likelihood functions are leptokurtotic, meaning they asymptote to zero more gradually than Gaussians ("heavy" tails, Figure 1a), their product (i.e., the posterior) is no longer Gaussian. The shape of the posterior, which may be skewed or multi-modal, is then determined by the size of the conflict relative to the variance and tail heaviness of the likelihoods (intersection of joint-likelihood cross with cues-consistent line in Figure 1c). When the likelihoods' tails are heavier than Gaussians' and the conflict is small, the product is nearly Gaussian, so behavior is nearly identical to weighted averaging of the cue values. When the conflict is large, robustness can occur when one arm of the joint-likelihood cross intersects the cues-consistent line (Figure 1c). We will refer to this as the *heavy-tailed likelihood model.*

Experiment 1

The first experiment was designed to determine how slant estimates from binocular disparity and the texture gradient are combined as a function of conflict size. We chose disparity and texture for three reasons. First, they have been extensively studied, so their geometry and the manner in which the visual system measures and uses them are reasonably well understood. Second, the parameters that affect the reliability of disparity (primarily viewing distance and secondarily slant magnitude) differ from those that affect texture reliability (primarily texture regularity and slant magnitude), and that allowed us to manipulate relative reliability as we wished. Third, when the conflict is small, disparity and texture slant cues are combined in an optimal weighted average consistent with the Gaussian-likelihood model (Hillis et al., 2004; Knill & Saunders, 2003), so a departure from the Gaussian prediction should be evident.

If robust behavior occurs and one cue is rejected, we wanted to know if there is a pattern in the choice of which cue is rejected. One hypothesis is that texture is always rejected because the underlying generative model that assumes that textures are homogenous and isotropic can be false (e.g., irregular surface markings on an abstract painting), while disparity is more likely to be a veridical indicator of slant (Knill, 2007). A second hypothesis is that the cue with lower statistical reliability will be rejected because it is more likely to have been corrupted by noise and therefore to be an invalid indicator of slant.

To distinguish these hypotheses, we needed to use stimuli with particular relative reliabilities. For the purposes of determining what stimuli to use, we assumed the likelihoods were Gaussian, and defined the relative reliability ratio as, r_D:r_T where, r_i is $1/\sigma_i^2$ is the variance and D and T are disparity and texture, respectively. To distinguish weighted averaging from robustness, the reliabilities cannot differ too greatly. For example, if $r_D \ll r_T$, the Gaussian and heavy-tailed models would predict the same behavior: perceived slant close to the texture-specified slant, S_T. The relative reliabilities cannot be too similar either. For example, if $r_D \simeq r_T$ and the observer behaved robustly by always choosing disparity, we would not be able to determine the reason: it could be that robustness occurs because texture is always ignored with large conflicts, or it could be that the observer always chooses the more reliable stimulus, and the disparity stimulus was just slightly more reliable. Therefore, we chose reliability ratios of 3:1 (disparity more reliable) and 1:3 (texture more reliable) because those values made it possible to discriminate the possibilities under consideration. For each observer, we found a range of conflict sizes and viewing distances that from single-cue measurements would yield approximately those relative reliability ratios.

Methods

Observers

Six observers participated, including the first author (S1). Five were unaware of the experimental hypotheses (S2–S6). All had normal or corrected-to-normal visual acuity and stereoacuity according to standard clinical tests.

Apparatus

All stimuli were displayed on a custom stereoscope with two mirrors and two CRTs (one for each eye; Backus, Banks, van Ee, & Crowell, 1999). Each mirror and CRT was attached to an arm that rotated about a vertical axis that was co-linear with the eye's rotation axis. The lines of sight from the eyes to the centers of the CRTs were always perpendicular to the CRT surface. To get the eyes' rotation axes in the appropriate position, head position relative to the apparatus was adjusted precisely using a sighting device and a bite bar (Hillis & Banks, 2001). With this arrangement, the mapping between the stimulus and retinas remained the same even as we altered the vergence-specified distance.

We used anti-aliasing to specify dot and line positions to sub-pixel accuracy. We spatially calibrated each CRT to eliminate distortions in the images (Backus et al., 1999). The optical distance between the center of rotation of each eye and the CRT face was 39 cm. A diffusing filter was placed just in front of each CRT to make the pixels invisible and to blur the images and thereby minimize the effects of the blur- and accommodation-specified distance to the display (Watt, Akeley, Ernst, & Banks, 2005).

Stimuli

Stimuli were virtual planes slanted about a vertical axis (i.e., tilt = 0°). We independently manipulated two slant cues: disparity and the texture gradient. There were single-cue and two-cue conditions. In the former, we presented either disparity-only or texture-only stimuli. In the disparity-only condition, the stimuli were viewed binocularly; in the texture-only condition, the irregular or regular texture stimuli were viewed monocularly.

The texture stimulus was the perspective projection of planar patches textured with Voronoi patterns (de Berg, van Kreveld, Overmars, & Schwarzkopf, 2000; Figure 2b). On a fronto-parallel plane, a regular grid of points was defined. To create an irregular texture stimulus (Figure 2b), we perturbed each point on the dot grid by random amounts horizontally and vertically according to a uniform distribution from −0.5° to 0.5°. Voronoi cells defined by these points were then computed. The resulting textured plane was rotated by an amount equal to the texture-defined slant. For the irregular textures, the width co-varied with slant but had a random component so outline shape was not a reliable cue to slant. The visible portion of the irregular texture stimulus was elliptical with a fixed height of 22°. To create a regular texture stimulus (Figure 2c), no random perturbation was applied to the points on the grid nor to its width. It had a fixed height of 48°.

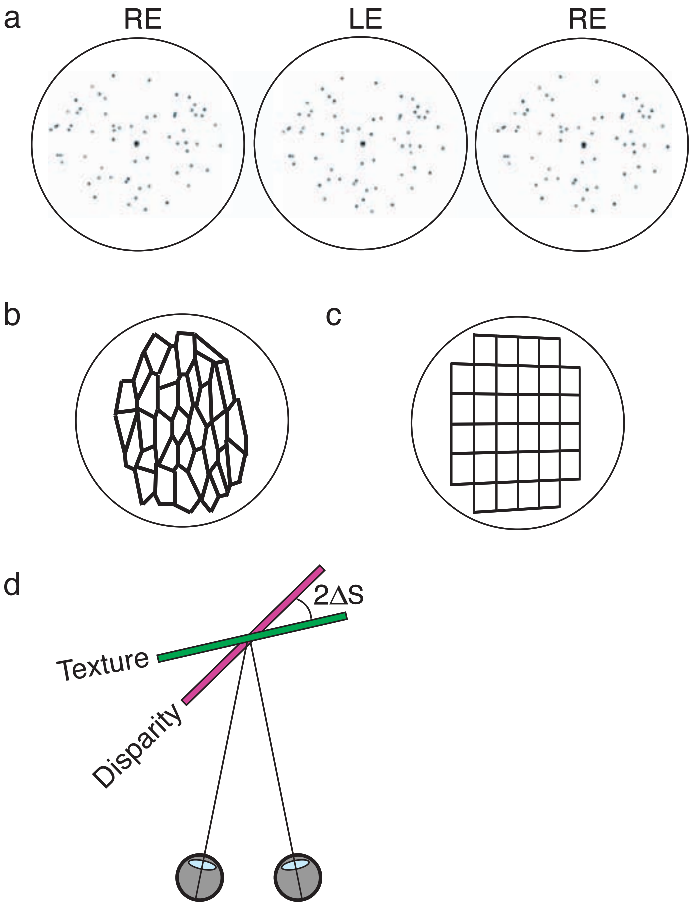

Figure 2. Experimental stimuli. a) Disparity-only stimulus. Cross-fuse the left and center panels to see the random-dot stimulus in 3D. Or divergently fuse the center and right panels. b) Irregular texture stimulus (monocular). c) Regular texture stimulus (monocular). d) Plan view of the conflict stimuli. The pink line represents the disparity-specified slant and the green line the texture-specified slant. They differ by 2Δ*S*.

We presented irregular and regular textures for theoretical reasons. The probability of encountering a non-homogenous, anisotropic texture is presumably less with quite regular textures than with irregular ones. For this reason, the tails of the texture likelihood function are likely to be less heavy when the stimulus is regular as opposed to irregular. If so, observers who behave robustly should be more likely to rely on the texture cue when the stimulus is regular. We will refer to the two conditions as the irregular-texture and regular-texture conditions.

The disparity slant cue was the difference between left- and right-eye projections, calculated separately for each observer's inter-ocular distance. In the two-cue conditions, the stimulus was defined by the Voronoi textures described above. In the disparity-only conditions, it was defined by sparse random dots (Figure 2a). In the latter case, the stimulus contained 75 dots, on average, whose positions were randomly drawn from a uniform distribution in the fronto-parallel plane. After the stimulus was rotated to the disparity-specified slant, the monocular dot gradient was consistent with the disparity-specified slant. Dot density was chosen randomly on each trial from the range of 0.125 to 0.25 dots/deg^2. The visible stimulus height was 20° and width was 25° ± 5°. The virtual viewing distance, specified by vergence and pattern of vertical disparities, ranged from 15 to 157 cm (see below). We verified that the random-dot stimulus did not contain useful monocular slant information by presenting it to all observers with one eye occluded. In all cases, the monocular random-dot stimulus was not a reliable cue to slant: observers were either not able to make slant discriminations at all, or performed much worse than with binocular information. Thus observers were quite unlikely to have used monocular slant information in the disparity-only stimulus.

In the two-cue conditions, the irregular or regular texture stimuli were viewed binocularly. Disparity and texture were either consistent ("no-conflict", $S_D = S_T$) or in conflict ("cue-conflict" $S_D \neq S_T$; Figure 2d). In the no-conflict cases, homogeneous textured surfaces were projected directly to the two eyes. In cue-conflict cases, we first calculated a perspective projection of the texture with slant S_T at the Cyclopean eye and then found the intersections of rays through this Cyclopean projection with a surface patch at the disparity-specified slant S_D. The markings on this latter surface were then projected to the left and right eyes to form the two monocular images.

Task

The task was two-interval, forced-choice slant discrimination. Each trial had two stimulus presentations of 1500 ms, separated by a blank of 750 ms. Observers indicated whether the first or second stimulus had more signed slant (i.e., right side farther away). No feedback was provided. A fixation point was presented during and in-between stimuli to aid accurate and stable fixation.

Single-cue conditions, procedure, and analysis

To estimate the single-cue reliabilities, we measured texture-only and disparity-only discrimination thresholds at base slants of 75, 60, 52.5, 45, 30, 15, and 0°. On each trial, one interval contained the standard stimulus at one of the base slants *S*, and the other contained the comparison stimulus at $S \pm \delta S$. The interval order was random. We used adaptive staircase procedures to vary δS. The procedures had two reversal rules—2-down/1-up and 1-down/2-up—to distribute points along the psychometric function. Two to four staircases were employed for each psychometric function, corresponding to an average of 124 trials per function. In each session, one base slant and one cue were presented with two interleaved staircases. Sessions were conducted in random order. The virtual viewing distance was constant within a given experimental session.

To estimate each observer's single-cue variances, we first fit each set of psychometric data with a cumulative Gaussian using a maximum-likelihood criterion and variable lapse rate up to 5% (Wichmann & Hill, 2001a, 2001b). Because we used a two-interval psychophysical procedure, we divided the standard deviations of the resulting functions by $\sqrt{2}$ to produce estimates of the standard deviations of the underlying likelihood functions, σ_D and σ_T (Green & Swets, 1974). These values are the just-noticeable differences (JNDs), the slant differences that are correctly discriminated ~84% of the time. Figure 3a shows single-cue JNDs as a function of slant. In general, texture JNDs decreased (improved) as the absolute value of slant increased; JNDs with the regular texture were generally lower than with the irregular texture. Disparity JNDs generally decreased as the absolute value of slant increased, but sometimes increased at large slants. As one would expect, disparity JNDs increased with viewing distance.

To infer JNDs for all possible slants, plots of measured JND versus base slant were fit with smooth interpolating functions. Those functions, which are described in the caption to Figure 3a, provided good fits to the data.

Two-cue conditions, procedure, and analysis

To achieve reliability ratios of 3:1 and 1:3, we had to present different stimuli to different observers because their single-cue JNDs, and therefore their relative cue reliabilities, differed. For each observer, we computed a relative reliability surface to represent the changing reliabilities for different stimuli and conflicts (Figure 3b). (This novel design may prove useful in other contexts because it helps create the situations in which model predictions can be most readily distinguished.) We then found contours on that surface that contained stimuli with the desired reliability ratios (white lines in Figure 3b). Along each contour, there is a range of conflict sizes with the same reliability ratio (white dots in Figure 3b). We chose nine conflicts: $|2\Delta S| = 0, 6, 11, 22, 45, 60, 90$, and $120°$ where $\Delta S = S_D - S_T$. We found the points S such that $S_D = S + \Delta S$ and $S_T = S - \Delta S$ along each of the two contours for each ΔS for each observer; this yielded 18 conflict stimuli. We could not always find stimuli that produced the desired reliability ratios for the full range of conflict sizes, so we did not test all 18 conditions in all observers.

The virtual viewing distance was varied to maintain the same relative reliabilities for each conflict regardless of the texture regularity. The two-cue stimuli with irregular textures and the disparity-only stimuli were always viewed at 114 cm. The texture-only stimuli were viewed monocularly and thus did not have a specified viewing distance. We used the known linear relationship between ln(HSR) (a quantification of the horizontal disparity gradient, see Figure 3) and discrimination threshold (Hillis et al., 2004) to estimate the viewing distance needed for the two-cue stimuli with regular textures to have the same

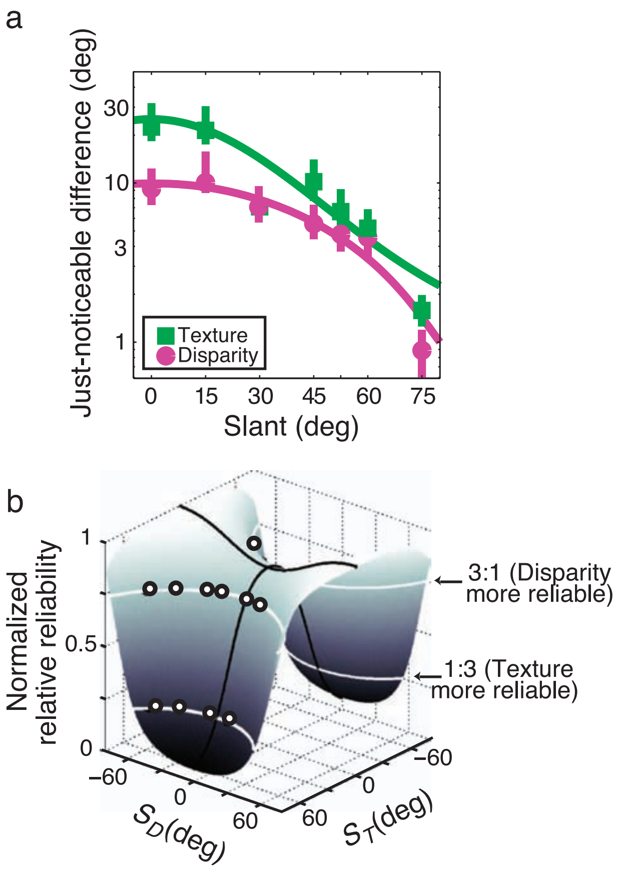

Figure 3. Just-noticeable differences and relative reliabilities. a) JNDs as a function of slant for representative observer S1. Pink circles and green squares represent disparity and texture, respectively. We fit curves to the disparity JNDs by first converting the data to horizontal-size ratios (HSR), defined as α_L/α_R where α_L and α_R are the horizontal angles subtended by a surface patch in the left and right eyes, respectively. The data were then fitted with a line in log space with two free parameters such that JND = ω exp(βHSR) (Hillis et al., 2004). The texture JNDs were fit with a scaled Gaussian with two free parameters: JND = θN(0, ϕ). b) Relative reliability surface for same observer. Normalized relative reliability is plotted as a function of the disparity- and texture-specified slants, where normalized reliability is $r_D/(r_D + r_T)$ for $r_i = 1/\sigma_i^2$. The normalized reliability is a re-writing of the reliability ratio, r_D:r_T that allows us to plot it between 0 and 1. Using the JNDs from a, we computed r_D and r_T (see text). For each possible combination of disparity and texture slants, we computed the normalized relative reliability for the various combinations of disparity- and texture-specified slants. The surface shows the cue conflicts for which disparity is most reliable (peaks) and for which texture is most reliable (troughs). Intersections of the surface and planes parallel to the floor create contours of constant relative reliability. The white lines show two such contours of interest: the desired relative reliability ratios of 3:1 (top line, disparity more reliable) and 1:3 (bottom line, texture more reliable). Points along these contours have constant reliability ratios, but varying conflict sizes. The white circles indicate the conflict conditions used in the experiment. This procedure was done separately for each observer.

relative reliability ratios. The disparity-only stimuli were re-measured at these distances to confirm that the relative reliability ratios were maintained reasonably close to desired values of 3:1 and 1:3.

The procedure in the two-cue conditions was the same as in the single-cue conditions except that a conflict stimulus and a no-conflict stimulus were presented in random order on each trial. The irregular and regular textures were presented in different experimental sessions for the two-cue conditions. Two staircases were employed for each psychometric function, for an average of 82 trials per function. One base slant was presented in each session with two interleaved staircases.

At the end of the experiment, observers were asked what the large-conflict stimuli looked like. All of them reported stable percepts, not bi-stable ones as described by van Ee et al. (2002). Some said that some of the stimuli occasionally looked “weird” or “distorted”.

We again fit the two-cue data from each condition with a cumulative Gaussian and variable lapse rate of up to 5% using a maximum-likelihood criterion (Wichmann & Hill, 2001a, 2001b). The no-conflict stimulus that on average had the same perceived slant as the conflict stimulus (i.e., the mean of the fitted function) was our estimate of the point of subjective equality, or PSE.

Results

Figure 4 shows the two-cue predictions and results for two representative observers. Most observers behaved similarly to observer S1 (left column). Observer S3 (right column) behaved quite differently. The rows represent different combinations of reliability ratio and texture regularity. The black circles in each panel represent the conflict stimuli and the yellow triangles represent the no-conflict stimuli that matched the conflict stimuli in perceived slant; yellow lines connect the conflict/no-conflict pairs that had equal perceived slants. We refer to the directions of those lines as cue-combination directions. The red lines are the matches predicted by the Gaussian model; they are consistent with weighted averaging. The blue lines represent the matches expected from the heavy-tailed model (discussed below); they are consistent with weighted averaging at small conflicts and with robustness at large conflicts.

The data show that when the cue conflict was small, the two observers exhibited weighted averaging whether the texture was regular or not (i.e., the yellow and red lines have similar and oblique directions). This is consistent with both the Gaussian and heavy-tailed models, the details of which will be described below. As the conflict size increased, observers tended to exhibit robust behavior (i.e., the yellow and blue lines have similar directions which are either horizontal or vertical), which is consistent with the heavy-tailed model, but inconsistent with the Gaussian model. When these observers exhibited robustness with a given texture regularity, their percept was always consistent with one cue and never with the other, regardless of the relative reliability ratio; it only depended on the conflict size and the texture regularity.

Figure 5 summarizes the data from all of the observers. It plots the cue-combination direction for the irregularly and regularly textured stimuli as a function of normalized conflict size (see caption). When the irregular texture was presented (left side) five of the six observers (S1, S2, S4, S5, S6) became robust at large conflict sizes choosing disparity. Observer S3 (and one data point from S4) showed the opposite type of robustness: she matched slant according to texture. When the regular texture was presented (right side of Figure 5), all six observers matched slants as if they were using only the texture signal. This dramatic change in behavior must be due to the change in texture regularity because the base slants, conflict sizes, and reliability ratios were the same with the irregular and regular textures.

The fact that observers often followed texture exclusively is evidence against the Knill’s hypothesis that texture is always rejected in favor of disparity (choose-disparity model; Knill, 2007). The fact that the rejected cue was not always the less reliable cue is evidence against the hypothesis that robust observers discount the less reliable cue (choose-reliable model).

The transition from weighted averaging to robustness occurred at a wide variety of conflict sizes across observers and conditions. We wondered if the critical conflict size was a particular value relative to the JNDs for the two cues. In Figure 5, there is a general trend toward robustness with greater conflicts, but there does not appear to be a particular value at which the transition from weighted averaging to robustness occurs. We conclude that the conflict size at which the transition to robustness occurs is not systematic; rather it varies from one stimulus condition to another and from one observer to another.

For the Gaussian model the likelihood distributions were, of course, Gaussian. For the heavy-tailed model, we could have used a variety of distributions (e.g., Student’s T, Pareto, Gaussian plus uniform), but we chose mixtures of Gaussians because they are simple and compatible with Knill’s suggestion of using mixture models to capture the contributions associated with multiple sources of information. The likelihoods were:

$$p(X|S) = \lambda_1 \mathrm{N}(\mu_1,\sigma_1) + (1-\lambda_1)\mathrm{N}(\mu_2,\sigma_2), \qquad (2)$$

where λ ranges between 0 and 1 and is the probability of the primary Gaussian distribution with primary variance σ_1^2 and mean μ, and $1 - \lambda$ is the probability of the secondary Gaussian with secondary variance σ_2^2 and the same mean μ. We assumed that the secondary Gaussian distribution produced the heavy tails: i.e., $\sigma_2 > \sigma_1$. We also assumed that JNDs from the single-cue measurements

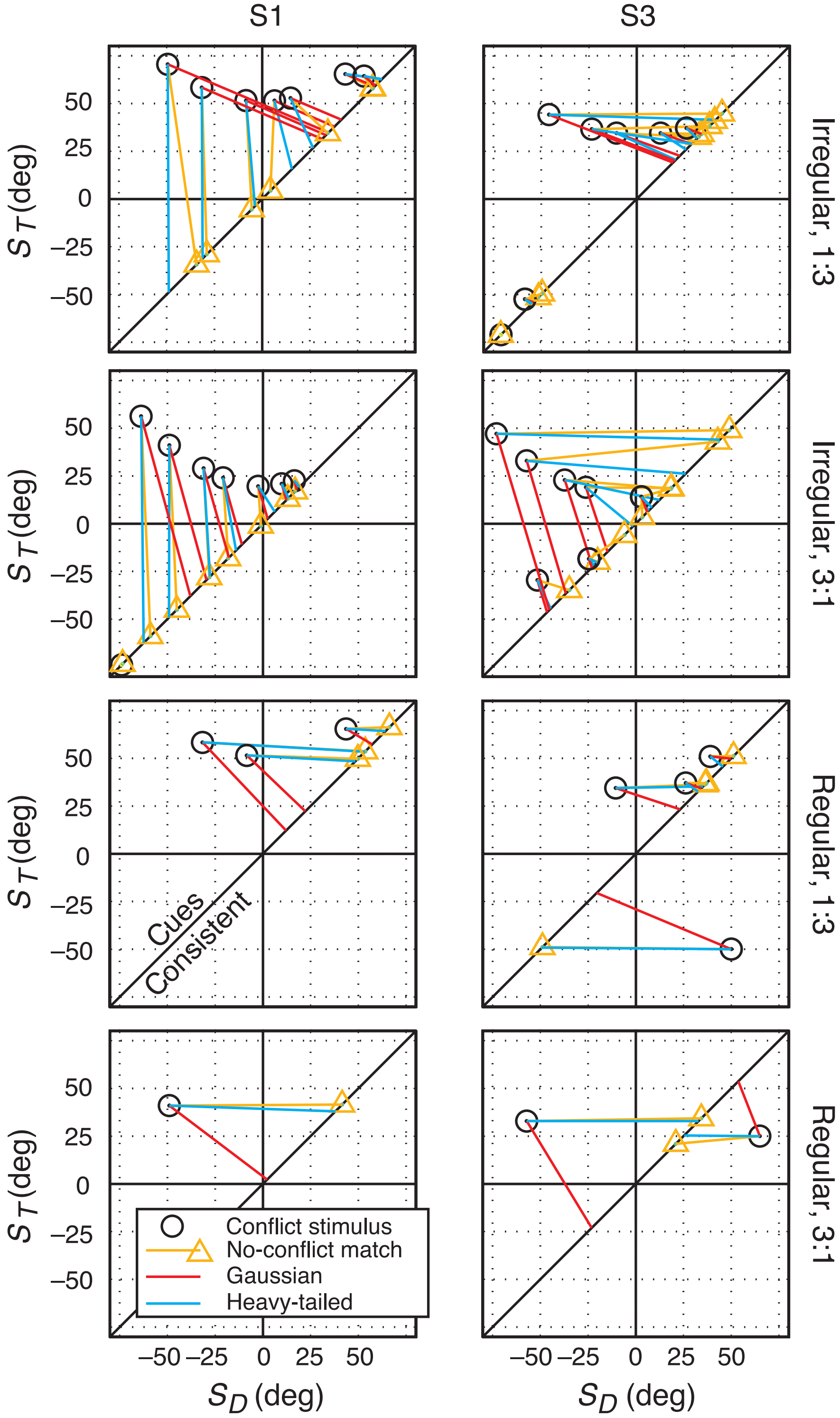

Figure 4. Predictions and results of Experiment 1 for observer S1 (left column) and S3 (right column). Each row represents a different condition. Rows 1 and 3 represent conditions in which the relative reliability ratio was 1:3 (texture more reliable) and rows 2 and 4 represent conditions in which the ratio was 3:1 (disparity more reliable). The first two rows represent data when the texture was irregular and the last two represent data when the texture was regular. Note that we were able to achieve the same reliability ratios with irregular and regular textures with the same slants by adjusting the distance to the disparity stimulus. Each panel plots disparity-specified and texture-specified slant on the horizontal and vertical axes, respectively. Black circles represent the conflict stimuli and yellow triangles the no-conflict stimuli that had the same perceived slant as the conflict stimuli; the yellow lines connect the appropriate stimulus pairs. The red lines connect the conflict stimuli with the no-conflict stimuli that the Gaussian likelihood model predicts to have the same perceived slant. The blue lines connect the conflict stimuli with the no-conflict stimuli that the heavy-tailed model predicts to have the same perceived slant.

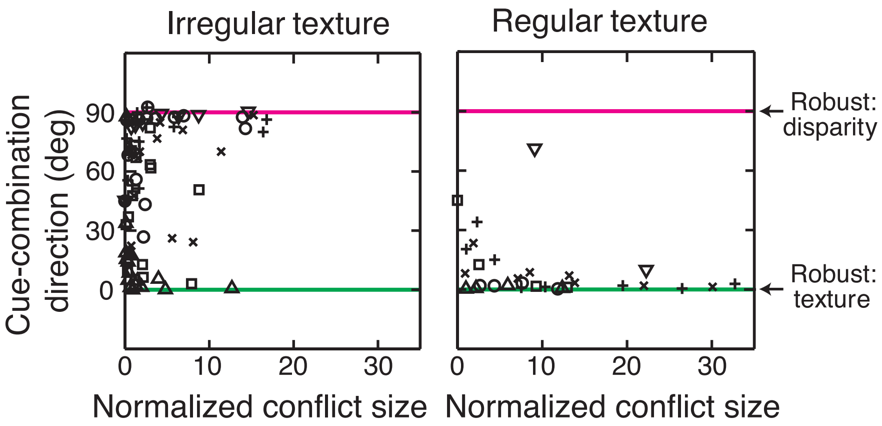

Figure 5. Cue-combination directions in Experiment 1 for all observers. The left and right panels are for the irregular and regular textures, respectively. Different symbols represent the data from the six observers (S1: o, S2: x, S3: Δ, S4: □, S5: +, S6: ▽). The horizontal axis is normalized conflict size (conflict size divided by the pooled standard deviation, $\sqrt{\sigma_D^2 + \sigma_T^2}$, for the two cues, akin to d-prime). The cue-combination direction is the angle between the data vector in Figure 4 and the horizontal axis. The robust choosing disparity prediction is at 0° (pink horizontal line). The robust choosing texture prediction is at 90° (green horizontal line). The predictions for the Gaussian model would be horizontal lines at 60° for cases in which disparity was more reliable (3:1) and 30° for cases in which texture was more reliable (1:3).

(Figure 2a) provided an estimate of σ_1, which is reasonable for $\sigma_2 \gg \sigma_1$ and $\lambda \geq 0.5$. Thus, we set σ_1 based on the single-cue discrimination thresholds leaving σ_2 as a free parameter. We observed that λ and σ_2 had similar effects on the fits to the data (decreasing λ was very similar to increasing σ_2), so we fixed λ at 0.5 leaving σ_2 as the only free parameters (i.e., $p(X\,|\,S) = 0.5\mathrm{N}(\mu_1,\sigma_1) + 0.5\mathrm{N}(\mu_2,\sigma_2)$). There were, therefore, two free parameters per observer: σ_{D2} for disparity and σ_{T2} for texture. σ_{T2} was estimated separately for the irregular and regular textures. The joint likelihood in the heavy-tailed model is the product:

$$p(X_d|S)p(X_t|S) = \frac{1}{8\pi}\left[\frac{e^{\frac{-0.5(x_D-\mu_D)^2}{\sigma_{D1}^2}}}{\sigma_{D1}} + \frac{e^{\frac{-0.5(x_D-\mu_D)^2}{\sigma_{D2}^2}}}{\sigma_{D2}}\right]\left[\frac{e^{\frac{-0.5(x_T-\mu_T)^2}{\sigma_{T1}^2}}}{\sigma_{T1}} + \frac{e^{\frac{-0.5(x_T-\mu_T)^2}{\sigma_{T2}^2}}}{\sigma_{T2}}\right]. \tag{3}$$

Figures 1b and 1c schematize the workings of the Gaussian and heavy-tailed likelihood models, assuming a uniform prior. We designed the task such that observers made discriminations along the cues-consistent line. Thus, the model considered the marginals of the posterior along that line. As the conflict size in the two-cue stimulus increases, the peak in the Gaussian model remains along a line connecting the white and blue dots in Figure 1b. The direction of this line is determined entirely by the relative reliabilities, not the conflict size, so the Gaussian model produces weighted averaging for all conflict sizes. The heavy-tailed model behaves differently. As the conflict size in the two-cue stimulus increases, the peak gradually moves from the weighted average to align either horizontally or vertically with one of the cues (a horizontal line, not shown, connecting the white and blue dots in Figure 1c). Thus, the models make quite different predictions of observers' matches at large conflicts.

We found the values of the two free parameters of the heavy-tailed model that yielded the best fit to the data by maximizing the log likelihood of the data given the model. On each trial, we assumed a uniform prior and computed the posteriors for the first and second interval using Equation 3. The model's decision was calculated numerically as the probability that one random draw from the comparison posterior was greater than a random draw from the standard posterior (Mamassian & Landy, 1998). By fitting all the raw data, we found the parameter values that yielded predicted PSEs and JNDs that were most similar to the observed values. This was done separately for each observer.

Table 1 shows the estimated secondary variances for all observers. Larger secondary variances correspond to heavier tails. We are interested in relative tail heaviness, which is indexed by the ratio of secondary variances. In most observers, the estimated texture likelihoods had heavier tails than the disparity likelihoods when the stimulus had an irregular texture (i.e., $\sigma_{T2} > \sigma_{D2}$). When the stimulus had a regular texture, the texture likelihoods had less heavy tails than the disparity likelihoods (i.e., $\sigma_{D2} > \sigma_{T2}$). This occurred because most observers chose disparity when robust with irregular texture and chose texture with regular texture and consequently, the tail heaviness in the modeling changed to capture this change

			σ_{T2}		σ_{T2}	
	σ_{D2}	*SD*	Irregular	*SD*	Regular	*SD*
S1	21.01	1.12	54.74	16.31	4.92	0.31
S2	33.85	16.83	59.52	25.52	7.45	2.46
S3	70.10	13.88	22.29	6.15	6.09	0.66
S4	74.20	20.10	64.56	16.55	6.75	7.59
S5	36.59	17.58	70.85	31.50	8.58	3.73
S6	10.58	4.56	16.28	4.61	10.68	2.93

Table 1. Experiment 1 estimated secondary variances for each observer. Standard deviations on the parameter estimates were calculated from bootstrapping the data 50 times.

in behavior. In sum, the estimated secondary variances are consistent with the ignored cue having heavier tails.

Throughout this paper, we assume the Gaussian model when referring to the relative reliability ratio (3:1 or 1:3) which was used to determine the initial stimulus conditions. It is important to note that under the heavy-tailed likelihood model, the relative reliability ratio can change. Under the heavy-tailed and Gaussian models, the relative reliability ratio is r_D:r_T where, r_i is 1/JND$_i^2$ and JND$_i$ is the slant difference that is correctly discriminated ~84% of the time. The difference in the two models is the shape of the psychometric function: cumulative Gaussian for the Gaussian model and a more complex function for the heavy-tailed model. In the cases in which observers were robust choosing the less reliable cue according to the Gaussian model, this cue was always more reliable under the heavy-tailed model. Said another way, under the heavy-tailed model, the observer always chooses the more reliable cue (i.e., the cue with the smaller JND). Thus, in the framework of the heavy-tailed model, observers' behavior was statistically optimal.

To determine the model that provided the best account of the results, we also calculated the sum of squared error between the data and seven models: the Gaussian likelihood model (degrees of freedom = 0), the heavy-tailed likelihood model (df = 2), a "choose-disparity" model in which the observer was robust and always chose disparity (df = 0), a "choose-texture" model in which the observer was robust and always chose texture (df = 0), a "choose-reliable" model in which the observer was robust and always chose the cue that was most reliable under the Gaussian model (df = 0), a coin-flipping model (random choice on every trial; df = 0), and psychometric fitting for each condition (df = 20–48, depending on the number of conditions per observer). The coin-flipping model and psychometric-fitting models provide estimates respectively of the lower and upper bounds on goodness of fit. Any model that provides a poorer fit than coin flipping is producing choices that are less consistent with the data than random choices.

The goodness of fit for each model is shown in Figure 6. The heavy-tailed model was a much better predictor of the data than the Gaussian, choose-disparity, choose-texture, and choose-reliable models. Indeed, it fit the data nearly as well as the psychometric-fitting model. The Gaussian model, which is the standard in the cue-combination literature, provided a rather poor fit to the data: it cannot account for robust behavior, so its performance declined dramatically with increasing conflict size causing it to provide a poorer fit than even the random, coin-flipping model. Thus, the heavy-tailed model provides the best account of the data. Observer S6 completed fewer conditions than the others, so there was less data to constrain his model fits. We also examined whether the heavy-tailed model provides the best account once it is penalized for the additional free parameters it has relative to the Gaussian, choose-disparity, choose-texture, and choose-reliable models. We computed Bayesian Information Criteria (BIC) (Burnham & Anderson, 2002) for all observers and all models and found decisive evidence for the heavy-tailed model in all cases.

The data that went into Figure 6 include all conflict sizes. As we said earlier, we expected the Gaussian and heavy-tailed models to perform similarly at small conflicts and quite dissimilarly at large ones. To see if this was the case, we divided the data into small-conflict (≤11°) and large-conflict (>11°) groups. The Gaussian and heavy-tailed models both performed well when the conflict was small (across subjects: 0.72 ± 0.13 and 0.78 ± 0.06, respectively, in goodness-of-fit units where the psychometric fits were 1 and the coin-flipping model was 0) because they can both exhibit weighted-averaging behavior at small conflicts. At large conflicts, the Gaussian model performed very poorly, even more poorly than coin flipping (−0.39 ± 0.14), but the heavy-tailed model still performed reasonably well (0.56 ± 0.16). This analysis therefore confirms that the Gaussian and heavy-tailed models can capture weighted-averaging behavior at small conflicts, but that only the heavy-tailed model can capture behavior at large conflicts where observers exhibited robustness. It is also interesting to note that the choose-reliable model performed poorly at small conflicts (0.094 ± 0.19) and even worse for large conflicts (−0.42 ± 0.25), indicating that robust observers did not simply choose the more reliable cue.

Just-noticeable differences (JNDs)

We were interested in examining just-noticeable differences (JNDs) because they are relevant to hypotheses about bi-modality and fusion. Figure 7 plots normalized JNDs as a function of normalized conflict size. The estimates are noisy because of the difficulty in estimating the slope of the psychometric function without a larger number of trials. The figure reveals that there was a slight tendency for JNDs to increase with conflict size ($\rho = 0.42$, $p < 10^{-5}$).

van Ee et al. (2002) reported that disparity-texture conflict stimuli with small conflicts generally yield one

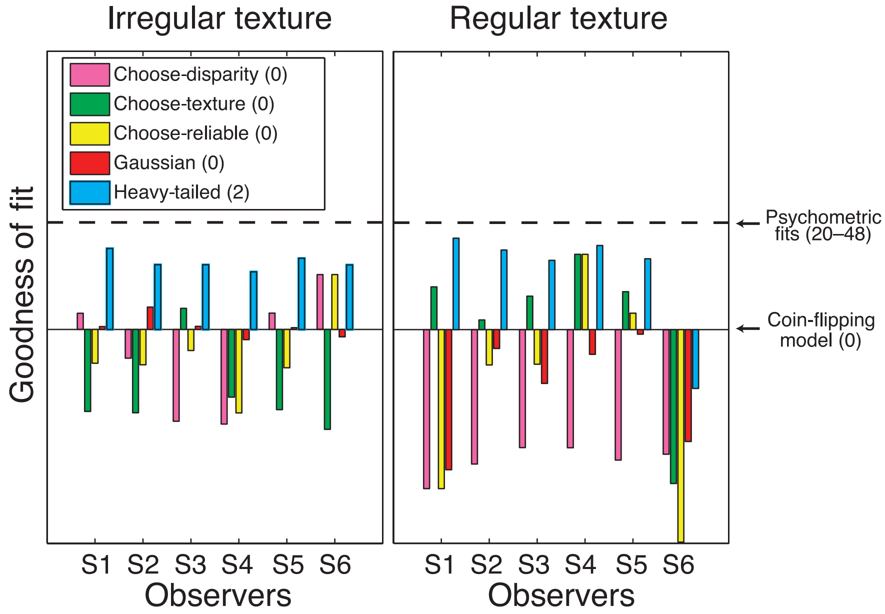

Figure 6. Analysis of the results from Experiment 1. Goodness of fit is plotted for each observer and model. The left panel shows the outcome with irregular textures, and the right panel the outcome with regular textures. Goodness of fit was calculated by measuring the sum of squared error between the data and predictions for each of the seven models. Those errors were then normalized, separately for each observer, with the psychometric-fitting and coin-flipping models providing the upper and lower bounds, respectively. The number of free parameters for each model is indicated in parentheses.

consistent slant percept while stimuli with large conflicts often yield bi-stable percepts, one similar to the disparity-specified slant and one to the texture-specified slant. If this occurred in our experiment, we should have observed very large JNDs at large conflicts as observers randomly switched from one percept to the other. Specifically, switching from one cue to another would yield plateaus in the psychometric functions. The widths of the plateaus would be proportional to the conflict size, yielding very large estimated JNDs. We did not see any evidence for such plateaus and, as Figure 7 shows, no large increase in JNDs with increasing conflict size. Thus, we did not observe the strong relationship between conflict size and JND that one expects with bi-stability. This is consistent with our observers' reports that they perceived only one slant whether the conflict was small or large.

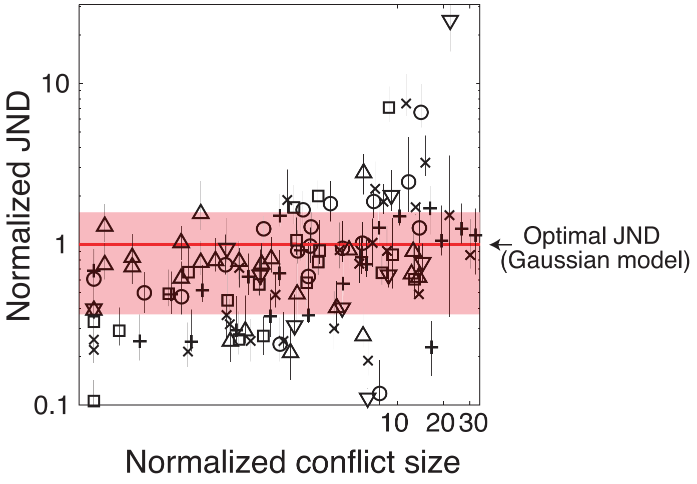

Figure 7. JND data from Experiment 1 for all observers. Normalized JND is plotted as a function of normalized conflict size. We normalized each JND by dividing by the optimal JND for that conflict stimulus. The optimal JND was calculated using the corresponding single-cue JNDs determined from the relative reliability surface and the Gaussian-likelihood model: $\sqrt{\sigma^2_{D1}\sigma^2_{T1}/(\sigma^2_{D1}+\sigma^2_{T1})}$. Error bars are 95% confidence intervals. The abscissa is on a log scale; normalized conflicts of 0 are plotted at 0.1. Different symbols represent the data from the six observers (S1: o, S2: x, S3: Δ, S4: □, S5: +, S6: ▽). The red horizontal line indicates a normalized JND of 1, the point of optimal performance with the Gaussian-likelihood model; the mean region of 95% confidence is shown in light red.

Experiment 2

In the first experiment, we examined how two sensory signals—disparity-specified slant and texture-specified slant—are combined to form coherent percepts. Those combined percepts were similar to a weighted average when the two signals specified nearly the same slant and were similar to robust estimation when they specified quite different slants. Both of these effects are consistent with a probabilistic model with heavy-tailed likelihood functions. We next asked if observers retain conscious access to the constituent signals and if that access depends on whether the combination behavior is similar to weighted averaging or robust estimation.

Hillis, Ernst, Banks, and Landy (2002) claimed that access is lost to disparity and texture slant signals when they are combined. In their experiment, they presented two kinds of slant stimuli: cues-consistent and cues-inconsistent. In a three-interval oddity task, they presented on every trial either two consistent stimuli and one inconsistent stimulus, or one consistent and two inconsistent stimuli. Observers identified the odd one among the three. There were many cues-inconsistent stimuli that could not be discriminated from a cues-consistent stimulus even though the disparity signals and/or the texture signals in the inconsistent and consistent stimuli were very different. Interestingly, these same signals that were indiscriminable in two-cue stimuli were readily discriminated when presented in isolation. Hillis and colleagues concluded that the visual system combines disparity and texture signals into a single percept and, in the process, loses conscious access to the individual cue values. We will refer to this sort of behavior as *complete fusion*. Hillis and colleagues used the same paradigm to examine the combination of cues from different senses. They found that observers could usually make oddity discriminations among two-cue, visual-haptic stimuli based on the separate visual and haptic stimulus values. This means that conscious access was maintained to the visual and haptic signals. We will refer to this behavior as *no fusion*.

Although the disparity-texture data were compelling, there were two potential shortcomings with the Hillis et al. (2002) conclusion that complete fusion occurs with disparity and texture. First, the task required that the observer remember three stimuli before responding. Indeed, the observer conceivably had to remember three signals—disparity, texture, and combined—in each of the three stimuli for a total of nine signals. This requirement may have encouraged the observer to concentrate on one signal—the combined one—to reduce the memory load. If that were the case, the finding might have more to do with the properties of short-term memory than with the properties of perception. Second, Hillis and colleagues did not measure whether the observers' percepts were similar to weighted averaging or to robustness, so they may not have presented cues-inconsistent stimuli that had sufficiently large conflicts to lead to robustness.

Experiment 2 was designed to determine if complete fusion is still observed when the memory load is greatly reduced and when the disparity-texture conflict stimulus is perceived robustly. For each of the conflict conditions and stimulus reliabilities in Experiment 1, we determined whether the same observers exhibited complete fusion, no fusion, or something in-between. We did so by finding the slant of a disparity-only stimulus that matched the perceived slant of a disparity-texture conflict stimulus and by finding the slant of a texture-only stimulus that matched the perceived slant of the same disparity-texture conflict stimulus. We assumed that if observers did not completely fuse, their matches would differ. For example, if they did not fuse at all, they would match the disparity-only stimulus to the disparity component of the conflict stimulus, and likewise for texture. If they partially fused, their disparity-only match would be closer to the disparity component of the conflict stimulus than to the texture component and their texture-only match would be closer to the texture component than to the disparity component of the conflict stimulus.

Methods

The observers and apparatus were the same as in Experiment 1 as were the stimuli, relative reliability ratios, and viewing distances. In the two-interval, forced-choice task, one interval contained a two-cue conflict stimulus, and the other contained one of two single-cue stimuli (disparity only or texture only). After the two intervals were presented, the observer indicated the one containing the greater perceived slant. No feedback was provided. The single- and two-cue stimuli were presented in random order, and the three types of single-cue conditions were randomly interleaved. We used a 1-up/1-down staircase procedure to vary the slant of the single-cue stimulus. As before, we fit the psychometric data with a cumulative Gaussian and used the mean of the fitted function as the estimate of the PSE: the slant of the single-cue stimulus that on average had the same perceived slant as the two-cue stimulus. Two to four staircases were run for each condition, corresponding to 53 trials per condition on average.

Results

Figure 8 shows the predictions and results. If the disparity and texture signals were completely fused, as reported by Hillis et al. (2002), the observer would have conscious access to only the combined slant estimate and not to the disparity- and texture-specified slants. As a consequence, the single-cue stimulus that matched the two-cue stimulus would have the same slant whether it was a disparity-only or texture-only stimulus. The predicted matches for complete fusion are thus along the cues-consistent line in Figure 8. On the other hand, if the disparity and texture signals were not fused at all, the observer would have access to both the disparity-specified and texture-specified slants. Consequently, the single-cue stimulus that would perceptually match the two-cue stimulus would have the same slant as the corresponding signal in the two-cue stimulus. Consider, for example, a two-cue stimulus with disparity and texture slants of −25° and 25° that has, let us say, a perceived slant of 0°. If

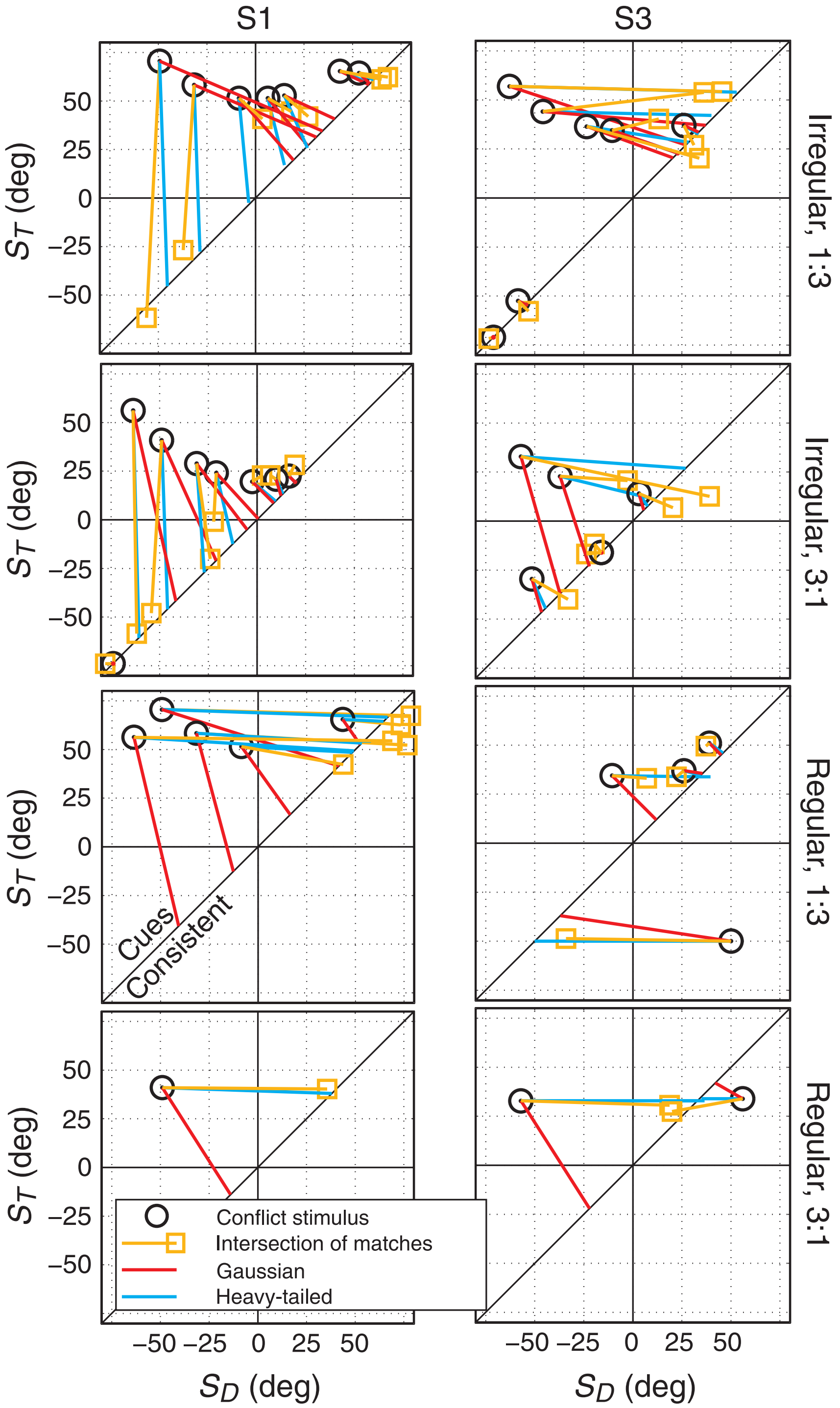

Figure 8. Predictions and results of Experiment 2 for observer S1 (left column) and S3 (right column). Rows 1 and 3 represent conditions in which the relative reliability ratio was 1:3 (texture more reliable) and rows 2 and 4 represent conditions in which the ratio was 3:1 (disparity more reliable). The first two rows represent data when the texture was irregular and the last two represent data when the texture was regular. Each panel plots disparity-specified and texture-specified slant on the horizontal and vertical axes, respectively. Black circles represent the two-cue, conflict stimuli. The single-cue stimuli that perceptually matched the two-cue stimuli can be visualized as horizontal (texture-only) and vertical (disparity-only) lines (not shown). Yellow squares represent the intersection of these two lines, and thus represent two settings at once. Yellow lines connect those matching stimuli. The red lines represent the predictions of the Gaussian model (see Discussion); the positions where those lines intersect the cues-consistent line represent the predicted matches if complete fusion occurred; matches consistent with partial fusion lie along the same lines, but closer to the two-cue conflict stimulus. The blue lines represent the predictions of the heavy-tailed model (see Discussion); matches consistent with complete fusion lie at the intersection of those lines and the cues-consistent line; matches consistent with partial fusion are on the same lines, but closer to the two-cue conflict stimulus.

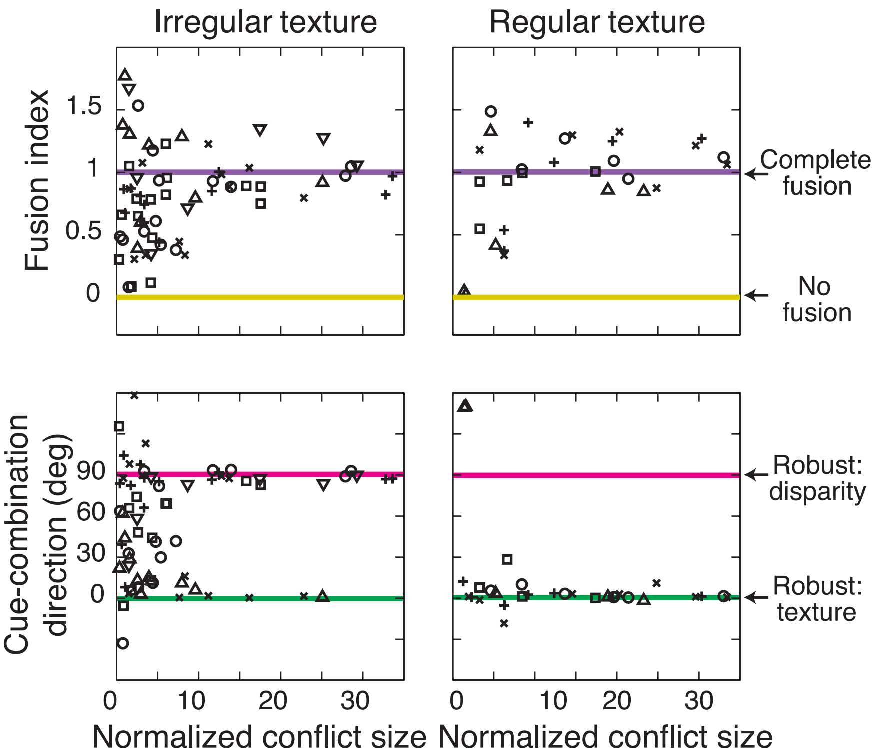

Figure 9. Summary of results in Experiment 2 for all observers. The upper row shows the observed amount of fusion relative to the no-fusion and complete-fusion predictions. The fusion index (see text) is plotted as a function of normalized conflict size. An index of 0 indicates no fusion (yellow horizontal line). An index of 1 indicates complete fusion (yellow purple line). The left and right columns are for the irregular and regular textures, respectively. Different symbols represent the data from the six observers (S1: o, S2: x, S3: Δ, S4: □, S5: +, S6: ▽). The second row shows cue-combination directions as a function of normalized conflict size. The robust choose-disparity prediction is 0° (pink horizontal line), indicating the match was determined entirely by the texture signal. The robust choose-texture prediction is 90° (green horizontal line), indicating the match was determined entirely by the disparity signal. The Gaussian predictions would be horizontal lines at 60° when disparity was more reliable (3:1, according to the Gaussian model) and at 30° when texture was more reliable (1:3).

complete fusion occurred, the observer would set the slants of both the disparity-only and texture-only stimuli to 0° to match the perceived slant of the two-cue stimulus (intersection at (0°, 0°)). If no fusion occurred, the slant of the matching disparity-only stimulus would be −25° and the slant of the matching texture-only stimulus would be 25° (intersection at (−25°, 25°)). The predicted match in Figure 8 would thus lie on top of the conflict stimulus.

The data in Figure 8 show that complete fusion was characteristic of essentially all the matches. Specifically, the yellow squares are always close to the cues-consistent line, which is the prediction for complete fusion. The *fusion vector* is the yellow line between the two-cue stimulus (black circle) and the two matching single-cue stimuli (yellow square). The red and blue lines represent the predicted fusion vectors for the Gaussian and heavy-tailed models, respectively. The vector direction for the Gaussian model is independent of conflict size while the direction for the heavy-tailed model rotates toward horizontal or vertical with increasing conflict size. The Gaussian and heavy-tailed predictions are thus similar when the conflict is small, but become quite dissimilar with increasing conflict size as robustness occurs and the vector predicted by the heavy-tailed model rotates toward horizontal or vertical.

The top row of Figure 9 shows the complete fusion result in another way. The *fusion index* is the length of the fusion vector divided by the distance from the two-cue conflict stimulus to the cues-consistent line along the vector. An index of 1 indicates complete fusion and an index of 0 no fusion. The indices in Figure 9 are spread more widely at small conflicts because the normalization involved in computing the fusion index has an increasingly small denominator with small conflict sizes and so the index becomes more sensitive to measurement noise. The figure shows quite clearly that fusion indices were centered around 1, indicative of complete fusion, for all conflict sizes. The fact that the data are quite consistent with complete fusion, particularly at large conflicts, is inconsistent with the finding of van Ee et al. (2002) that conflict stimuli yield bi-stable percepts.

The direction of the fusion vector indicates which cue(s) determined the match, and corresponds to the cue-combination direction in Experiment 1. When the vector

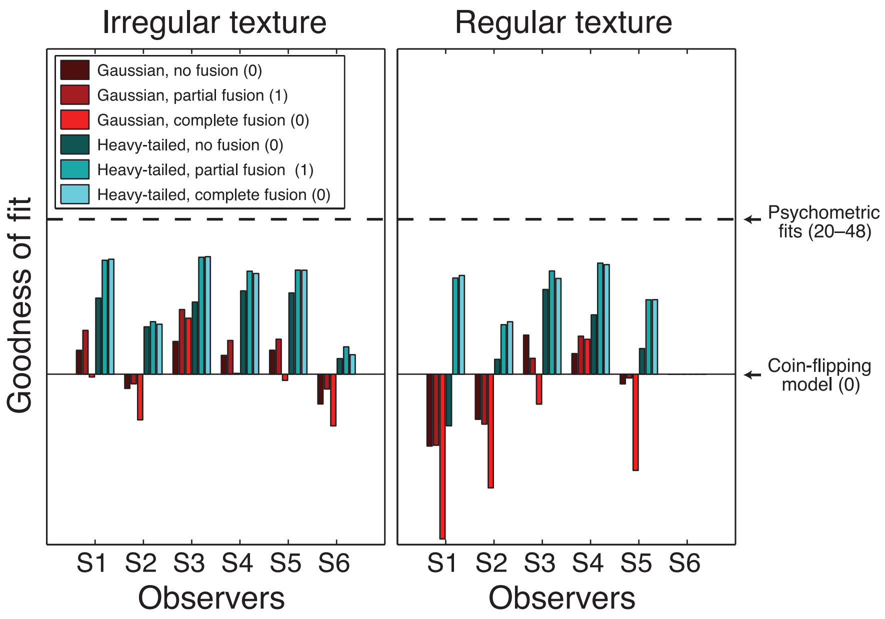

Figure 10. Relative goodness of fit for various models of Experiment 2. The abscissa represents the six observers. The ordinate represents the goodness of fit, computed as in Figure 6. Dark red, medium red, red, dark blue, blue, and light blue represent the goodness of fit respectively for the various models: Gaussian no-fusion, Gaussian partial-fusion, Gaussian complete-fusion, heavy-tailed no-fusion, heavy-tailed partial-fusion, and heavy-tailed complete-fusion. The number of free parameters is indicated in parentheses. The left panel shows the data when the texture was irregular and the right panel the data when it was regular. The dashed lines represent the fits for the psychometric and coin-flipping models, which respectively represent upper and lower bounds for goodness of fit.

is horizontal in Figure 8, the match was determined by only the texture component of the conflict stimulus. When it is vertical, the match was determined by the disparity component only. If the fusion vector is oblique, both cues contributed to the match, which is indicative of weighted averaging. The bottom row of Figure 9 shows that at small conflicts, the cue-combination directions were generally consistent with weighted averaging (although noise in the calculation of the normalized conflict size at small conflicts contributed to the spread of the data). At large conflicts, the cue-combination directions were either horizontal (0°) or vertical (90°) consistent with both matches being made according to the texture and disparity components of the two-cue stimulus, respectively.

We examined which model of likelihoods and fusion best accounted for the data in another way. In causal-inference models (Körding et al., 2007; Sato et al., 2007), cue fusion is linked to the probability of inferring one as opposed to two causes for two sensory measurements. We simulated two versions of the causal model: one with Gaussian likelihoods (Körding et al., 2007; Sato et al., 2007) and another with heavy-tailed likelihoods (see Supplement). Körding et al. (2007) used four free parameters, one called p_{common} that equals the prior probability of one cause p(C = 1), where C equals the number of causes, and three others. We simulated three variants of the model: no fusion (p_{common} = 0), complete fusion (p_{common} = 1), and partial fusion (p_{common} as a free parameter). The likelihood parameters (σ^2_{D1}, σ^2_{D2}, σ^2_{T1}, and σ^2_{T2}) were set by measurement and analysis in Experiment 1. The prior over slant was assumed to be uniform; we later verified that this assumption had no impact on our main conclusions. We then measured goodness of fit in the same fashion as we functions (df = 16–48 depending on the observer), coin flipping (df = 0), Gaussian, no fusion (df = 0, p_{common} did for Experiment 1. Eight models were tested: psychometric = 0), Gaussian, partial fusion (df = 1), Gaussian, complete fusion (df = 0, p_{common} = 1), heavy-tailed, no fusion (df = 0, p_{common} = 0), heavy-tailed, partial fusion (df = 1), and heavy-tailed, complete fusion (df = 0, p_{common} = 1). The goodness of fit was normalized such that the psychometric and coin-flipping models represented the upper and lower bounds, respectively. Figure 10 shows the results. The goodness of fit for the heavy-tailed likelihood was always better than for the Gaussian likelihood, consistent with Experiment 1. Among the heavy-tailed models, the goodness of fit for the complete-fusion model was also consistently greater than for no-fusion model and was nearly as great as for the partial-fusion model, which had a free parameter. The fits for the partial-fusion model were quite similar to the fits for the complete-fusion model, suggesting that adding p_{common} as a free parameter rather than fixing it at 1 is not necessary to account for these data. Indeed, the mean best-fitting p_{common} for the heavy-tailed, partial-fusion model was p_{common} = 0.88(±0.08), which is quite close to 1. Computing Bayesian Information Criteria (BIC) (Burnham & Anderson, 2002), which penalizes the

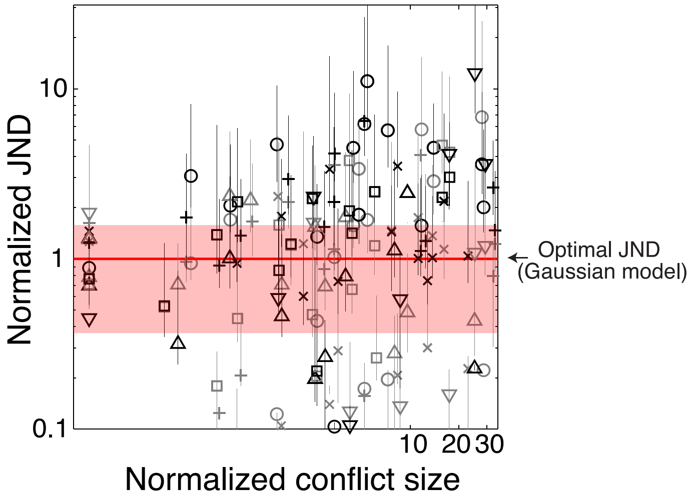

Figure 11. JND data from Experiment 2 for all observers. Normalized JND is plotted as a function of normalized conflict size, as in Figure 7. Gray symbols represent the data from disparity-only matches and black symbols data from texture-only matches.

partial-fusion models for their one free parameter, revealed decisive evidence for all observers in favor of the heavy-tailed, complete-fusion model. The results therefore suggest that observers assumed one cause, i.e., $p(C = 1) \approx 1$.

Can the causal-inference model account for our data? Yes, but only if it incorporates heavy-tailed likelihood functions and assumes *a priori* that there is only one cause.

Just-noticeable differences (JNDs)

Figure 11 plots normalized JNDs as a function of normalized conflict size. The figure reveals that there was a tendency for JNDs to increase with conflict size (ρ = 0.29 ($p < 0.005$) for disparity-only matches and ρ = 0.19 ($p < 0.05$) for texture-only matches). As we discussed in the analysis of JNDs in Experiment 1, the increase in JND is much less than one would expect if cue switching.

In summary, we found in Experiment 2 that conscious access to single-cue estimates does not accompany the robustness we observed in Experiment 1. Said another way, observers can have a robust percept in which they followed one signal and ignored the other and yet completely fuse the discrepant signals.

Discussion

Comparison to previous work

Several previous reports have examined cue-combination behavior with large signal discrepancies and found that observers seem to have conscious access to the individual signals. Nearly all of these reports involved signals in two sensory modalities (the visual-haptic experiment in Bresciani, Dammeier, & Ernst, 2006; Hillis et al., 2002; Körding et al., 2007; Roach, Heron, & McGraw, 2006; Sato et al., 2007; Shams, Ma, & Beierholm, 2005; Wallace et al., 2004). To our knowledge, only one has claimed that observers have conscious access to discrepant within-modality signals of the same environmental property (van Ee et al., 2002, and subsequent reports from that lab using the same paradigm). Hillis et al. (2002) did not observe such behavior and neither did we. Interestingly, Hafter and Carrier (1972) observed nearly complete fusion between inter-aural time difference and inter-aural intensity difference in a sound-localization task.

We cannot pinpoint the cause of the difference between the current observations and those of van Ee and colleagues, but their paradigm and ours differed in several ways, and many of the differences could plausibly have contributed. van Ee et al. (2002) explicitly instructed observers to attend separately to the two cues in their conflict stimulus. Indeed, they gave special instructions on how to access the disparity and texture signals: they told observers to assume either that a trapezoidal shape on the screen was a slanted square (in which case they were meant to indicate the texture slant) or a slanted trapezoid (in which case they were meant to indicate the disparity slant). We provided no such instructions. Instead, we assessed cue access by having observers make single-cue matches to two-cue stimuli. van Ee and colleagues told their observers that cue conflicts would be present and we did not. They asked observers during the course of the experiment if they were aware of cue conflicts; we did not ask until after the experiment was completed. Their stimulus duration was quite long (several seconds) and ours was brief (1.5 sec). Given the different time courses needed to achieve stable estimates of disparity and texture slant (van Ee & Erkelens, 1998), it is possible that our observers would have experienced bi-stable percepts if they had viewed the stimuli for longer periods of time. Given the significance of determining how cue combination occurs in the face of stimulus discrepancies, it will be important to resolve the cause of the discrepant outcomes in our experiments and those of van Ee and colleagues.

Knill (2007) also examined perceived slant when conflicting texture and disparity signals are presented. As the size of the conflict between the signals increased, his observers gradually down-weighted the texture signal becoming robust by choosing the disparity-specified slant. This behavior is consistent with heavy tails on the texture likelihood. In contrast, we showed in Experiment 1 for the regular textures, observers always chose the texture-specified slant when they transitioned to robustness, behavior consistent with heavy tails on the disparity likelihood. Indeed, our results showed that the relative tail heaviness depended on the regularity of the texture: tails were less heavy when the texture was regular.

As previously mentioned, many well-known illusions manifest robust behavior in which two different percepts are possible, but the viewer perceives only one at a time and never something in-between the two possibilities. The Necker Cube, for example, appears in one form or the other; the viewer does not perceive an average of the two interpretations and does not perceive both interpretations at the same time. In our terminology, the percepts are robust and completely fused. The binary nature of the percepts may be a manifestation of underlying heavy-tailed likelihoods for some or all of the depth signals present in those stimuli.

Plausibility of heavy-tailed distributions

The Gaussian distribution is nearly always used in the cue-combination literature to represent sensory likelihood distributions. The data reported here, however, are inconsistent with the Gaussian assumption. In the Bayesian framework, we had to use distributions with heavier tails than Gaussian to fit the data. Are heavy-tailed likelihood functions for surface orientation perception plausible?

Knill (2003) proposed that the texture likelihood might differ from Gaussian because for texture to be useful, the visual system must make one of at least two assumptions: that the texture on the viewed surface is isotropic and that it is homogenous. If the distributions associated with each assumption are Gaussian, their amalgamation can be expressed as a weighted sum of two Gaussians, each weighted by the probability that the corresponding assumption is true. The resulting distribution is a mixture of Gaussians, a distribution that is used frequently in engineering (Gelman et al., 2003; McLachlan & Peel, 2000).

Like Knill (2003), we assumed heavy tails for the texture likelihoods to account for the robust choose-texture behavior. In addition, we assumed heavy tails for the disparity likelihood to account for the robust choose-disparity behavior. Are heavy-tailed distributions for disparity plausible? We know that slant estimation from disparity is done in part from measurement of horizontal disparity and eye-position signals (Backus et al., 1999; Gårding, Porrill, Mayhew, & Frisby, 1995). This can be expressed by:

$$S = -\tan^{-1}(\ln(\mathrm{HSR})/\tilde{\mu} - \tan\gamma), \tag{4}$$

where HSR is the horizontal-size ratio (a measure of horizontal disparity), $\tilde{\mu}$ is the horizontal vergence of the eyes, and γ is the horizontal version of the eyes; the eye-position signals are measured via extra-retinal signals. Equation 4 is highly non-linear, so even if the noises associated with disparity and eye-position measurements were Gaussian distributed, the resulting distribution of slant estimates would not be. We simulated Equation 4 with Gaussian-distributed disparity and eye-position measurements and indeed observed heavy tails in the direction of greater slants. Thus, it seems reasonable to assume non-Gaussian distributions for the processes of estimating slant from disparity (see also Porrill, Frisby, Adams, & Buckley, 1999).

Perhaps, one could estimate the shapes of the underlying likelihood functions directly from the psychometric data in the single-cue measurements (Figure 3a) without having to use large conflicts at all. We examined this possibility and encountered two problems that make it unfeasible. First, response errors (the observer selecting the wrong response key), while infrequent, have a distressingly large effect on the estimation of the shape of the tails of the psychometric function. Unfortunately, it is precisely those parts of the psychometric function that are most important for distinguishing heavy-tailed and Gaussian likelihoods. Second, we conducted simulations of the ability to determine the shape of the underlying distribution even if no response errors occurred and found that the number of trials required in a psychophysical experiment was unfeasibly large. Thus, our evidence for the use of heavy-tailed distributions had to be obtained from slant matches when the discrepancy between disparity- and texture-specified slant was large.

Heavy-tailed likelihoods are advantageous in probabilistic inference because they protect the sensory system from errors due to measurements coming from different objects or parts of an object, when one or more of the measurements comes from a faulty sensor, or when the wrong generative model might have been used to interpret a measurement (e.g., assuming a trapezoidal texture instead of a square one). In other words, sensible behavior analogous to robust statistical estimation derives from the use of heavy-tailed likelihood distributions. This finding adds to accumulating evidence for heavy-tailed sensory distributions (Knill, 2007; Natarajan et al., 2009; Stocker & Simoncelli, 2006).

Does a heavy-tailed likelihood model predict bi-stable percepts? Generally it does not, but it can under some circumstances. The posterior would have to be bi-modal with roughly equally sized modes. This can occur at moderate conflict sizes (at both the small and large conflicts we considered, the product of the heavy-tailed distributions was unimodal). To observe bi-modal behavior, the decision rule would also have to sometimes choose one peak and sometimes choose the other. We did not observe clear evidence for a bi-modal posterior in any of our conditions.

Correlation between single-cue estimators

We examined whether our results could have been affected by an incorrect assumption about cue correlation. We made two common assumptions: First, we assumed that the noises associated with the measurements of disparity and texture were independent. This assumption is well justified in situations in which cues come from different senses, like vision and touch (Ernst & Banks, 2002). In the present study, the two cues are both visual, so they must share some sensory noise: they are both based on the same retinal image and therefore subject to the same Poisson statistics of the retinal quantum catch and the same neural noises in retinal processing. We do not know the relative magnitudes of the shared and independent noises. Second, we assumed that our single-cue stimuli isolated the cues of disparity and texture sufficiently to allow us to measure the variances of the texture and disparity estimators separately. This assumption seems valid for the texture-only stimulus because it was presented monocularly and minimized focus cues (Watt et al., 2005). The argument is less strong for the disparity-only stimulus because the stimulus has a texture cue to slant even if the cue is quite weak. We minimized the cue by using sparse random-dot textures (Hillis et al., 2004) and conducted a monocular control experiment to rule out reliable use of the monocular slant information in these stimuli. Thus, it seems quite unlikely that the measurement of the disparity estimator were contaminated by monocular cues.

If the noises associated with the measurements of disparity- and texture-specified slants are in fact correlated, we can examine how the correlation would thus affect the predictions of the Gaussian likelihood model. Oruç, Maloney, and Landy (2003) describe how to correct the variance when two cues are correlated with correlation coefficient ρ: to correct σ_{D1}^2, one divides it by $1 - \rho\sigma_{D1}/\sigma_{T1}$. The desired relative reliability ratios r_T:r_D in our experiments were 3:1 and 1:3. When corrected for correlation, these ratios become $[3 - \rho\sqrt{3}]$:$[1 - \rho\sqrt{3}]$ and $[1 - \rho\sqrt{3}]$:$[3 - \rho\sqrt{3}]$, respectively. In the 3:1 condition, disparity should have been more reliable than texture. As ρ increases in this condition, disparity becomes even more reliable relative to texture. Likewise, in the 1:3 condition, texture should have been more reliable than disparity. As ρ increases in this condition, texture becomes even more reliable relative to disparity. Thus, if the two cues were correlated, the actual reliabilities would not be as close to each other as we had estimated. This causes the predictions of the Gaussian likelihood model to be closer to the prediction of robustness choosing the more reliable cue. But such behavior is not consistent with much of our data because we often observed robustness in which the observer chose the less reliable cue. We conclude that an undetected correlation between the sensory noises would not substantively change the interpretation of our main findings.

Models of sensory fusion

When combining signals, an optimal observer should take into account the statistics of those signals, and the probability that the two measurements come from the same environmental source. Such an observer may exhibit complete, partial, or no fusion depending on the probability of one or two sources. Three models for sensory combination in this vein have been proposed recently: the coupling-prior model (Bresciani et al., 2006; Ernst, 2005; Roach et al., 2006), the causal-inference model (Körding et al., 2007; Sato et al., 2007), and the causal-selection model (Natarajan et al., 2009). The coupling prior represents the probability that two signals co-occur. The coupling-prior model makes predictions consistent with many aspects of cue combination, and consistent with the causal inference model (Körding et al., 2007).

The causal models determine how to combine sensory cues using two models of the causal structure—one cause (C = 1) or two (C = 2)—along with the prior probability of one cause, p(C = 1). The causal-inference model describes the percept as a weighted sum of the complete-fusion and no-fusion resultants. The prior p(C = 1) is a constant for any two sensory estimators. Thus, as conflict size increases, the causal-inference model makes a gradual transition from complete fusion through partial fusion to no fusion. The causal-selection model is similar except that it first selects C*, the number of causes that is the most probable, and then subsequent inference is based on C* alone. The causal models thus instantiate the simple idea that if there is likely to be a common cause, cues are combined and complete fusion occurs, and if there is not likely to be a common cause, they are segregated and no fusion occurs. The models' behavior is generally consistent with behavioral data in multi-sensory studies (e.g., Wallace et al., 2004). But the models are in an important sense not consistent with the data from our experiments. They link the departure from cue combination in part to the attribution of two causes determined by the conflict between the cues. We found instead that two causes were never perceived for our stimuli; observers fused completely no matter what the conflict size was. The causal models also link cue combination, manifest as weighted averaging, to the inference of one cause, and link cue segregation, which is similar to robustness, to the inference of two causes. Our data do not support a linkage between the transition from weighted averaging to robustness and the number of inferred causes. Perhaps the models are most readily applicable to cue combination between senses where the amount of cue fusion depends strongly on conflict size (Hillis et al., 2002; Wallace et al., 2004). Of course, the causal models could be modified to be consistent with our data by assuming p(C = 1) = 1 and by assuming that the underlying likelihoods are heavy tailed (Natarajan et al., 2009).

Acknowledgments

This research was supported by National Institutes of Health Grant R01 EY012851 and National Science Foundation Grant BCS-0117701. Thanks to Johannes

Burge, Mikhail Katkov, Marc Ernst, Michael Landy, Alan Stocker, Simon Watt, and Richard Zemel for useful comments. Part of this manuscript appeared in Girshick (2007).

Commercial relationships: none.

Corresponding author: Ahna R. Girshick.
Email: ahna@cns.nyu.edu.
Address: 6 Washington Place, New York, NY 10003, USA.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14,* 257–262. [PubMed] [Article]

Arnold, R. D., & Binford, T. O. (1980). Geometric constraints in stereo vision. *SPIE: Image Processing for Missile Guidance, 238,* 281–292.

Backus, B. T., Banks, M. S., van Ee, R., & Crowell, J. A. (1999). Horizontal and vertical disparity, eye position, and stereoscopic slant perception. *Vision Research, 39,* 1143–1170. [PubMed]

Banks, M. S., & Backus, B. T. (1998). Extra-retinal and perspective cues cause the small range of the induced effect. *Vision Research, 38,* 187–194. [PubMed]

Box, G. E. P., & Tiao, G. C. (1992). *Bayesian inference in statistical analysis*. Wiley Classics Library edition. New York: John Wiley & Sons.

Bresciani, J. P., Dammeier, F., & Ernst, M. O. (2006). Vision and touch are automatically integrated for the perception of sequences of events. *Journal of Vision, 6*(5):2, 554–564, http://journalofvision.org/6/5/2/, doi:10.1167/6.5.2. [PubMed] [Article]

Bülthoff, H. H., & Mallot, H. A. (1988). Integration of depth modules: Stereo and shading. *Journal of the Optical Society of America A, Optics and Image Science, 5,* 1749–1758. [PubMed]

Burnham, K. P., & Anderson, D. (2002). *Model selection and multi-model inference: A practical-theoretical approach* (2nd ed.). New York: Springer.

de Berg, M., van Kreveld, M., Overmars, M., & Schwarzkopf, O. (2000). *Computational geometry: Algorithms and applications* (2nd ed.). New York: Springer-Verlag.

Ernst, M. O. (2005). A Bayesian view on multimodal cue integration. In G. Knoblich, I. M. Thornton, M. Grosjean, & M. Shiffrar (Eds.), *Perception of the human body from the inside out* (pp. 105–131). New York: Oxford University Press.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415,* 429–433. [PubMed]

Gårding, J., Porrill, J., Mayhew, J. E. W., & Frisby, J. P. (1995). Stereopsis, vertical disparity and relief transformations. *Vision Research, 35,* 703–722. [PubMed]

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall.

Gepshtein, S., & Banks, M. S. (2003). Viewing geometry determines how vision and haptics combine in size perception. *Current Biology, 13,* 6, 483–488. [PubMed] [Article]

Ghahramani, Z., Wolpert, D. M., & Jordan, M. I. (1997). Computational models of sensorimotor integration. In P.G. Morasso & V. Sanguineti (Eds.), *Self-organization, computational maps and motor control*. Amsterdam: Elsevier Press.

Girshick, A. R. (2007). *Probabilistic integration of sensory information for 3D visual surface slant perception*. Ph.D. thesis, Berkeley: University of California.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics*. New York: Robert E. Krieger.

Hafter, E. R., & Carrier, S. C. (1972). Binaural interaction in low-frequency stimuli: The inability to trade time and intensity completely. *Journal of the Acoustical Society of America, 51,* 1852–1862. [PubMed]

Hay, J., & Pick, H. L., Jr. (1966). Visual and proprioceptive adaptation to optical displacement of the visual stimulus. *Journal of Experimental Psychology, 71,* 150–158. [PubMed]

Hillis, J. M., & Banks, M. S. (2001). Are corresponding points fixed? *Vision Research, 41,* 2457–2473. [PubMed]

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science, 298,* 1627–1630. [PubMed]

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of Vision, 4*(12):1, 967–992, http://journalofvision.org/4/12/1/, doi:10.1167/4.12.1. [PubMed] [Article]

Howard, I. P., & Templeton, W. B. (1966). *Human spatial orientation*. New York: Wiley.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Research, 39,* 3621–3629. [PubMed]

Knill, D. C. (2003). Mixture models and the probabilistic structure of depth cues. *Vision Research, 43,* 831–854. [PubMed]

Knill, D. C. (2007). Robust cue integration: A Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant. *Journal of Vision, 7*(7):5, 1–24, http://journalofvision.org/7/7/5/, doi:10.1167/7.7.5. [PubMed] [Article]

Knill, D. C., Kersten, D., & Yuille, A. (1996). A Bayesian formulation of visual perception. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference* (pp. 1–21). New York: Cambridge University Press.

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research, 43,* 2539–2558. [PubMed]

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE, 2,* e943. [PubMed] [Article]

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. J. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research, 35,* 389–412. [PubMed]

Mamassian, P., & Landy, M. S. (1998). Observer biases in the 3D interpretation of line drawings. *Vision Research, 38,* 2817–2832. [PubMed]

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York: John Wiley & Sons.

Natarajan, R., Murray, I., Shams, L., & Zemel, R. (2009). Characterizing response behavior in multisensory perception with conflicting cues. In *Advances in Neural Information Processing Systems 21*. Cambridge, MA: MIT Press.

Necker, L. A. (1832). Observations on some remarkable phenomena seen in Switzerland, and an optical phenomenon which occurs on viewing of a crystal or geometrical solid. *Philosophical Magazine, 1,* 329–337.

Ogle, K. N. (1938). Induced size effect: I. A new phenomenon in binocular space-perception associated with the relative sizes of the images of the two eyes. *Archives of Ophthalmology, 20,* 604–623.

Oruç, I., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Research, 43,* 2451–2468. [PubMed]

Pick, H. L., Warren, D. H., & Hay, J. C. (1969). Sensory conflict in judgments of spatial direction. *Perception & Psychophysics, 6,* 203–205.

Porrill, J., Frisby, J. P., Adams, W. J., & Buckley, D. (1999). Robust and optimal use of information in stereo vision. *Nature, 397,* 63–66. [PubMed]

Roach, N. W., Heron, J., & McGraw, P. V. (2006). Resolving multisensory conflict: A strategy for balancing the costs and benefits of audio-visual integration. *Proceedings of the Royal Society of London B: Biological Sciences, 273,* 2159–2168. [PubMed] [Article]

Rock, L., & Victor, J. (1964). Vision and touch: An experimentally created conflict between the two senses. *Science, 143,* 594–596. [PubMed]

Sato, Y., Toyoizumi, T., & Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism aftereffect: Identification of common sources of audiovisual stimuli. *Neural Computation, 19,* 3335–3355. [PubMed]

Shams, L., Ma, W. J., & Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *Neuroreport, 16,* 1923–1927. [PubMed]

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 9,* 578–585. [PubMed]

van Ee, R., & Erkelens, C. J. (1998). Temporal aspects of stereoscopic slant estimation: An evaluation and extension of Howard and Kaneko's theory. *Vision Research, 38,* 3871–3882. [PubMed]

van Ee, R., van Dam, L. C. J., & Erkelens, C. J. (2002). Bi-stability in perceived slant when binocular disparity and monocular perspective specify different slants. *Journal of Vision, 2*(9):2, 597–607, http://journalofvision.org/2/9/2/, doi:10.1167/2.9.2. [PubMed] [Article]

Wade, N. J., & Hughes, P. (1999). Fooling the eyes: Trompe l'œil and reverse perspective. *Perception, 28,* 1115–1119. [PubMed]

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research, 158,* 252–258. [PubMed]

Watt, S. J., Akeley, K., Ernst, M. O., & Banks, M. S. (2005). Focus cues affect perceived depth. *Journal of Vision, 5*(10):7, 834–862, http://journalofvision.org/5/10/7/, doi:10.1167/5.10.7. [PubMed] [Article]

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics, 63,* 1293–1313. [PubMed] [Article]

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics, 63,* 1314–1329. [PubMed] [Article]

Yuille, A. L., & Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference* (pp. 123–161). New York: Cambridge University Press.