PNAS
www.pnas.org

Supplementary Information for

Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence

Rachel N. Denison*[1,2], William T. Adler*[2], Marisa Carrasco[1,2], Wei Ji Ma[1,2]

[1]Department of Psychology, [2]Center for Neural Science,
New York University, New York, NY

Corresponding author:
Rachel Denison
Department of Psychology and Center for Neural Science
New York University
rachel.denison@nyu.edu

This PDF file includes:
Supplementary Text
Extended Materials and Methods
Figures S1–S7
Tables S1–S4
References for SI

* Equal author contribution

www.pnas.org/cgi/doi/10.1073/pnas.1717720115

Supplementary Text

S1 Theoretical motivations for the task

The goal of the current study was to test whether category and confidence decision rules account for attention-dependent uncertainty. Unlike the tasks used in previous studies, the task we used[1] can answer this question, because it has two properties: 1) Unlike the detection and coarse discrimination (e.g., ±45°) tasks used in most signal detection theory[2] (SDT) studies, the current task allows inference of absolute decision boundaries. 2) Unlike tasks with mirror-image categories (e.g., left vs. right discrimination), the current task creates an incentive to shift the category boundary when uncertainty changes.

S1.1 Inference of absolute decision boundaries

A decision rule can be thought of as a boundary defined on the observer's internal measurement space. Here we were interested in the absolute location of that boundary b. The "unified criterion" discussed previously also refers to an absolute boundary[3,4].

To infer absolute decision boundaries from behavioral data, the measurement axis must represent known feature values. The embedded category task has this property, because the measurement axis represents orientation. Making a category or confidence decision can be thought of as comparing the observed stimulus orientation to an internal reference orientation, which is the decision boundary. As experimenters, we know the means of the internal measurement distributions (specific orientations), so we can infer the absolute decision boundary on the orientation axis.

In SDT detection and coarse discrimination tasks, in contrast, absolute decision boundaries cannot be inferred, because the measurement axis represents values that we, as experimenters, do not know. In a detection task, the measurement value is thought of as the strength of the internal signal, or the "amount of evidence" that the external signal is present. In a coarse discrimination task, the measurement value is thought of as the amount of evidence for choice 1 (e.g., $-45°$) versus choice 2 (e.g., $+45°$). We don't know the means of the internal measurement distributions in real values; we don't even know what the units are. Consequently, the behavioral SDT measures d' (perceptual sensitivity) and c (criterion) are defined in a normalized space – d' and c are z-scored measures of the distance between the two internal category distributions and the location of the observer's decision boundary, respectively. So they are relative measures.

As a result, an absolute decision boundary b is unrecoverable from behavioral data. This fact can be shown mathematically. The standard formulae for d' and c are

$$d' = Z(H) - Z(F) \tag{S1}$$

$$c = -\frac{1}{2}(Z(H) + Z(F)), \tag{S2}$$

where Z is the inverse of the normal cumulative distribution function (i.e., z-score), H is the proportion of hits, and F is the proportion of false alarms. Note this formula gives c with respect to the unbiased criterion. If we let the mean of the noise distribution be 0 and the mean of the signal distribution be μ, then

$$d' = \frac{\mu}{\sigma} \tag{S3}$$

$$c = \frac{b - \frac{\mu}{2}}{\sigma}. \tag{S4}$$

Here we have two equations with three unknowns. Any combination of d' and c is therefore consistent with an infinite set of combinations of the μ, σ, and b parameters; thus b cannot be uniquely determined. The intuition here is that the SDT axis can be rescaled without changing d' and c (**Figure S1a**). The same issue applies not only to d' and c but to any other relative behavioral measure, such as hit rate or false alarm rate. Kontsevich et al.[5] raised this concern about Gorea and Sagi's[4] proposal of a unified criterion for simultaneously presented stimuli.

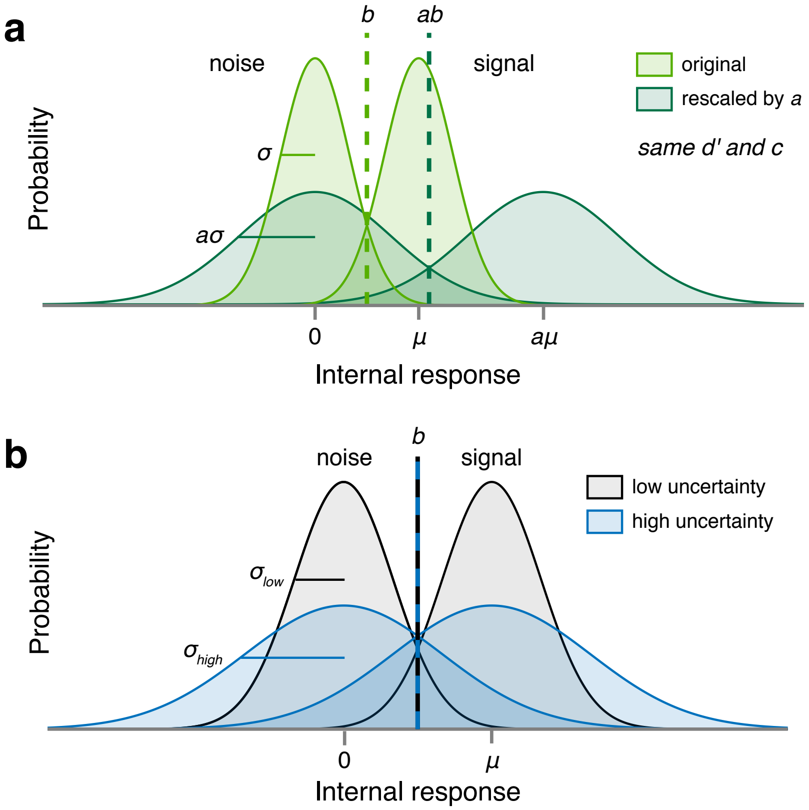

Figure S1: Methodological limitations in standard signal detection tasks. (**a**) Rescaling the SDT axis by a factor a yields the same values of d' and c, but with a different set of parameters (the original parameters rescaled by a). This is because d' and c are relative to the internal measurement distributions, not any absolute evidence metric. (**b**) In standard SDT tasks, when the means of the internal measurement distributions are symmetric about the optimal category boundary, changing the uncertainty does not change the optimal boundary. μ = mean, σ = standard deviation, b = decision boundary.

The non-uniqueness of SDT parameters creates a critical problem when asking whether b changes with attention. Attention could change μ, σ, or both properties of the internal measurement distributions[6–8]. Therefore, b cannot be compared, even in a relative fashion, across attention conditions; so fixed and flexible decision rules cannot be distinguished (**Figure 5**).

Note that in a left vs. right *fine* discrimination task, in which stimuli are drawn from orientation distributions with similar means, absolute decision boundaries can be inferred, again because the measurement axis represents orientation, and the measurement distribution means are known to the experimenter[9].

This argument relates to that by Aitchison et al.[10], who showed that to distinguish different models of confidence, two-dimensional data are required. They used features of two separate stimuli. Here, we used orientation and uncertainty.

S1.2 Incentive to shift the category boundary when uncertainty changes

Fine discrimination tasks have the first property, allowing inference of absolute decision boundaries. However, not every such task has the second property, an incentive to shift category boundaries.

In the embedded category task, the category distributions overlap in such a way that the optimal category boundaries shift when the uncertainty in the measurement distributions changes (**Figure 3a,b**). Therefore, observers have an incentive to shift their category decision rules when uncertainty changes, and we as experimenters are able to assess whether they do so.

In standard SDT tasks, in contrast, the optimal category boundary does not depend on uncertainty σ if the means of the internal measurement distributions remain symmetric about the boundary (**Figure S1b**). So if attention does not change the means, or changes them symmetrically (as in a discrimination task), then the optimal category boundary will not change. Observers therefore have no incentive to change their category decision rules when uncertainty changes, making it impossible to test whether the category boundary is fixed or flexible.

In summary, the embedded category task has two critical advantages over standard SDT tasks, which allow an unambiguous determination of whether and how perceptual decisions take uncertainty into account.

Extended Materials and Methods

S2 Experiment

S2.1 Observers

Twelve observers (7 female, 5 male), aged 18–25 years, participated in the experiment. These observers came from an original set of 28 observers who completed at least one session. The remaining observers did not complete the main experiment, either because they were not invited to continue following the pre-screening staircase sessions (15 observers, **Section S2.3.7**) or because they chose to stop participating before all sessions were completed (one observer). Observers received $10 per 40–60 minute session, plus a completion bonus of $25. The experiments were approved by the University Committee on Activities Involving Human Subjects of New York University. Informed consent was given by each observer before the experiment. All observers were naïve to the purpose of the experiment. No observers were fellow scientists.

S2.2 Apparatus and stimuli

2.2.1 Apparatus

Observers were seated in a dark room, at a viewing distance of 57 cm from the screen, with their chin in a chinrest. Stimuli were presented on a gamma-corrected 100 Hz, 21-inch display (Model Sony GDM-

5402). The display was connected to a 2010 iMac running OS X 10.6.8 using MATLAB (Mathworks) with Psychophysics Toolbox 3[11–13].

2.2.2 Stimuli

The background was mid-level gray (60 cd/m^2). Stimuli consisted of drifting Gabors with a spatial frequency of 0.8 cycles per degree, a speed of 6 cycles/s, a Gaussian envelope with a SD of 0.8 degrees of visual angle (dva), and a randomized starting phase. In category training, the stimuli were positioned at fixation, and the central fixation cross was a black "+" subtending 1.2 dva in diameter. In all other blocks, one stimulus was positioned in each of the four quadrants of the screen, at 45, 135, 225, and 315 degrees, 5 dva from fixation, and the fixation cross was a black "×" with each arm pointing to a quadrant. One or more of the arms turned white to provide a precue or response cue (Figure 1b). Stimulus contrast depended on the block type.

2.2.3 Categories

Stimulus orientations s_i were drawn from Gaussian distributions with means $\mu_1 = \mu_2 = 0°$, and standard deviations $\sigma_1 = 3°$ (category 1) and $\sigma_2 = 12°$ (category 2). Because the category distributions overlapped, maximum accuracy was ~80%.

2.2.4 Attention manipulation

During attention training and testing blocks, voluntary spatial attention was manipulated via a central precue presented at the start of the trial. A response cue at the end of the trial indicated which of the four stimuli to report. On each trial, each of the four stimuli was drawn from one of the two category distributions. Each stimulus was generated independently. In valid trials (66.7% of all trials), a single quadrant was precued and the response cue matched the precue. In invalid trials (16.7%), a single quadrant was precued and the response cue did not match the precue. Cue validity was therefore 80% when a single quadrant was precued. In neutral trials (16.7%), all four quadrants were precued, and the response cue pointed to one of the four quadrants with equal probability for each quadrant.

S2.3 Procedure

Each observer completed seven sessions. Because our behavioral task involved multiple components—orientation categorization, confidence reports, and attention—we trained observers on each component in a stepwise fashion, as described below.

The first two sessions ("staircase sessions") were used to pre-screen observers and find a stimulus contrast level that would achieve maximum separability in performance across the three attention conditions. Each staircase session consisted of 3 category training blocks and 3 category/attention testing-with-staircase blocks, in alternation. No confidence reports were collected in these sessions. The first category training block was preceded by a category demo, and the first category/attention testing-with-staircase block was preceded by a category/attention training block. Detailed instructions were provided in the first session. Most blocks consisted of sets of trials, in between which the observer was informed of their progress (e.g., "You have completed three quarters of Testing Block 2 of 3") and allowed to rest. The staircase sessions also served as practice on the categorization and attention components of the task, so that observers knew them well by the time they started the main experiment. During these sessions, stimulus contrast was 35% for training blocks, and varied during the testing-with-staircase blocks.

The final five sessions (“test sessions”) comprised the main experiment. Each test session consisted of 3 category training blocks and 3 confidence/attention testing blocks, in alternation. The first category training block was preceded by a category demo, and the first confidence/attention testing block was preceded by a confidence/attention training block. During these sessions, stimulus contrast was fixed to an observer-specific value in all blocks.

Combining all test sessions, 9 observers completed 15 confidence/attention testing blocks (2160 trials), 2 observers completed 14 testing blocks (2016 trials), and 1 observer completed 12 testing blocks (1728 trials). Accuracy on category training trials was 70.8% ± 4.0% (mean ± 1 SD) in staircase sessions and 71.9% ± 4.0% in test sessions, indicating that observers learned the category distributions well (recall that maximum accuracy on the task is ~80%).

2.3.1 Eye tracking

Eye tracking (Eyelink 1000) was used to monitor fixation online. In all blocks, trials were only initiated when the observer was fixating. In testing blocks, trials in which observers broke fixation due to blinks or eye movements were aborted and repeated later in the experiment.

2.3.2 Instructions

First staircase session. Before the first category training block, we provided observers with a printed graphic similar to Figure 1a, explained how the stimuli were generated from distributions, and explained the category training procedure. We also explained that trials would only proceed when the observer maintained fixation. Before the category/attention training block, we explained the attention task using an onscreen graphic that explained the cuing procedure and a printed graphic that illustrated cue validity. We also explained the requirement to maintain fixation from the precue until the response cue and the consequences of breaking fixation. Before the first category/attention testing-with-staircase block, we explained that the stimulus presentation time would be shorter and that the contrast of the stimuli would vary.

First test session. Before the confidence/attention training block, we explained two changes to the experiment. First, we told observers that they would be reporting category choice and confidence simultaneously. We provided a printed graphic similar to the buttons shown in Figure 1b, showing the eight buttons representing category choice and confidence level, the latter on a 4-point scale. The confidence levels were labeled as “very high,” “somewhat high,” “somewhat low,” and “very low.” All printed graphics were visible to observers throughout the experiment. Second, we told observers that contrast would be fixed (rather than variable) for the remainder of the experiment, in all blocks.

2.3.3 Category demo

We showed observers 25 randomly drawn exemplar stimuli from each category (50 exemplars in the first staircase session). Stimulus contrast was 35% in staircase sessions and observer-specific in test sessions.

2.3.4 Category training

To ensure that observers knew the stimulus distributions well, we gave them extensive category training with trial-to-trial correctness feedback and foveal stimulus presentation to reduce orientation uncertainty. Each trial proceeded as follows: Observers fixated on a central cross for 1 s. Category 1 or category 2 was selected with equal probability. The stimulus orientation was drawn from the corresponding stimulus distribution and displayed as a drifting Gabor. The stimulus appeared at fixation for 300 ms, replacing the fixation cross. Observers were asked to report category 1 or category 2 by pressing a button with their left or right index finger, respectively. Observers were able to respond immediately after the offset of the stimulus, at which point correctness feedback was displayed for 1.1 s, e.g., "You said Category 1. Correct!" The fixation cross then reappeared. In staircase sessions, the stimulus contrast was 35%. In test sessions, the contrast matched the observer-specific levels chosen for testing blocks, in order to minimize obvious changes between training and testing blocks. Each category training block had 2 sets of 36 trials (72 total). At the end of the block, observers were shown the percentage of trials that they had correctly categorized.

2.3.5 Category/attention training

To familiarize observers with the attention task before the testing-with-staircase blocks, they completed category/attention training. Observers performed the attention task, reporting only category choice. To prevent observers from forming a simple mapping of orientation measurement and attention condition onto the probability of category 1 (which might have biased behavior towards the Bayesian model), we withheld trial-to-trial feedback on this and all other types of attention blocks. The precue indicating which location(s) to attend to appeared for 300 ms, followed by a 300 ms period in which a standard fixation cross was shown. Then the four drifting Gabor stimuli were displayed for 300 ms. After another 300 ms period with a fixation cross, the response cue appeared, indicating which stimulus to report. The response cue remained on the screen until the observer pressed one of the two choice response buttons, with no time pressure. Observers were free to blink or rest briefly between trials, with a minimum intertrial interval of 800 ms. All attention conditions were randomly intermixed. The stimulus contrast was 35%, as in staircase session category training. The block had 36 trials in the first session and 30 trials in subsequent sessions. At the end of the block, observers were shown the percentage of trials they had correctly categorized.

2.3.6 Category/attention testing-with-staircase

The purpose of this block was to determine the stimulus contrast for each observer that would be used in the test sessions. The trial procedure was identical to that of category/attention training, except that stimulus presentation time was 80 ms (instead of 300 ms) and stimulus contrast varied. We used an adaptive staircase procedure to determine the stimulus contrast on each trial and estimate psychometric functions for performance accuracy as a function of log contrast. Separate staircases were used for valid, neutral, and invalid conditions. We used Luigi Acerbi's MATLAB (https://github.com/lacerbi/psybayes) implementation of the PSI method by Kontsevich and Tyler[14], extended to include the lapse rate[15]. The method generates a posterior distribution over three parameters of the psychometric function: threshold μ, slope σ, and lapse rate λ. On each trial, it selects a stimulus intensity that maximizes the expected information gain by completion of the trial. μ (log contrast units) ranged from -6.5 to 0 and had a Gaussian prior distribution with mean -2 and SD 1.2. $\log \sigma$ ranged from -3 to 0, and had a uniform prior distribution across the range. λ ranged from 0.15 (because the maximum accuracy in the task was slightly below $1 - 0.15$) to 0.5, and had a Beta prior distribution with shape parameters $\alpha = 20$ and $\beta = 39$. Each block had 4 sets of 36 trials (144 total). At the end of the block, observers were shown the percentage of trials that they had correctly categorized.

2.3.7 Observer pre-screening and contrast selection

Simulations we conducted before starting the study showed that without a sufficiently large noise (related to accuracy) difference between valid and invalid trials, our models would be indistinguishable. Therefore, we used a pre-screening process to select observers with a robust attention effect to participate in the main experiment. We also determined the stimulus contrast at which each observer's attention effect was maximal. This procedure increased the probability that uncertainty would depend on attention in the main experiment, which was critical for answering our central question about decision behavior. Note that the pre-screening procedure only concerned the overall accuracy difference between valid and invalid trials, which is independent of how attention affects the decision rule.

After each observer's final staircase session, we plotted and visually inspected the mean and SD of the posterior over the 3 (valid, neutral, and invalid) estimated psychometric functions (an example is shown in **Figure S7**). An observer was considered eligible for the remainder of the study if there existed a contrast that satisfied two conditions. 1) Invalid accuracy was above chance: The mean minus the SD of the posterior over invalid psychometric functions was above 0.5. 2) Valid accuracy was different from invalid accuracy: The mean minus the SD of the posterior over valid psychometric functions was greater than the mean plus 1 SD of the posterior over invalid psychometric functions. For example, note that there is a range of values in **Figure S7** for which the purple shading does not overlap with the chance line or with the green shading. Within the range of suitable contrasts, we selected the contrast for which the separation between valid, neutral, and invalid performance appeared to be maximal. Observers for which no suitable contrast could be found were not invited to participate in the main experiment. Selected contrasts ranged from 4% to 60% across observers.

2.3.8 Confidence/attention training

To familiarize observers with the button mappings for choice and confidence, they completed confidence/attention training. The trial procedure was identical to category/attention training, except observers reported their confidence on each trial in addition to their category choice. Observers were not instructed to use the full range of confidence reports, as that might have biased them away from reporting what felt most natural. Instead, they were simply asked to be "as accurate as possible in reporting their confidence" on each trial. Feedback about their choice and confidence report was presented for 1.2 s after each trial, e.g. "You said category 2 with HIGH confidence." The stimulus contrast was specific to each observer, based on the staircase sessions. There were 30 trials per block.

2.3.9 Confidence/attention testing

These were the main experimental blocks. The trial procedure (**Figure 1b**) was the same as in confidence/attention training blocks, but with no trial-to-trial feedback whatsoever. Each block had 4 sets of 36 trials (144 total). At the end of each block, observers were required to take a break of at least 30 s. During the break, they were shown the percentage of trials that they had correctly categorized. Observers were also shown a list of the top 10 block scores (across all observers, indicated by initials). This was intended to motivate observers to perform well, and to reassure them that their scores were normal, since it is rare to score above 75% on a block.

S3 Modeling

The modeling procedures were similar to those used by Adler and Ma[9]. Several modeling choices were adopted based on model comparisons performed for that study. These included: having orientation-dependent measurement noise; allowing all decision boundaries to be free parameters in the Bayesian model; including decision noise in the Bayesian model; and modeling three types of lapse rates.

S3.1 Measurement noise

We used free parameters to characterize σ, the standard deviation (SD) of orientation measurement noise, for all three attention conditions: $\sigma_{\text{valid}}, \sigma_{\text{neutral}}$, and σ_{invalid}.

We assumed additive orientation-dependent noise in the form of a rectified 2-cycle sinusoid, accounting for the finding that measurement noise is higher at noncardinal orientations[16]. For a given trial i, the measurement noise SD comes out to

$$\sigma_i = \sigma_{\text{attention condition}} + \psi \left| \sin \frac{\pi s}{90} \right|. \tag{S5}$$

The second term of this equation is a constant that depends on the stimulus orientation s, with ψ a free parameter that determines the degree of orientation dependence.

S3.2 Response probability

We coded all responses as $r \in \{1, 2, \ldots, 8\}$, with each value indicating category and confidence. A value of 1 mapped to high confidence category 1, and a value of 8 mapped to high confidence category 2, as in **Figure 1b**. The probability of a single trial i is equal to the probability mass of the internal measurement distribution $p(x \mid s_i) = \mathcal{N}(x; s_i, \sigma_i^2)$ in a range corresponding to the observer's response r_i. Because we only use a small range of orientations, we can safely approximate measurement noise as a normal distribution, rather than a von Mises distribution. We find the boundaries $(b_{r_i-1}(\sigma_i), b_{r_i}(\sigma_i))$ in measurement space, as defined by the fitting model m and parameters θ, and then compute the probability mass of the measurement distribution between the boundaries:

$$p_{m,\theta}(r_i \mid s_i, \sigma_i) = \int_{-b_{r_i}}^{-b_{r_i-1}} \mathcal{N}(x; s_i, \sigma_i^2)\, \mathrm{d}x + \int_{b_{r_i-1}}^{b_{r_i}} \mathcal{N}(x; s_i, \sigma_i^2)\, \mathrm{d}x, \tag{S6}$$

where $b_0 = 0^\circ$ and $b_8 = \infty^\circ$.

To obtain the log likelihood of the dataset, given a model with parameters θ, we compute the sum of the log probability for every trial i, where t is the total number of trials:

$$\log p(\text{data} \mid \theta) = \sum_{i=1}^{t} \log p(r_i \mid \theta) = \sum_{i=1}^{t} \log p_\theta(r_i \mid s_i, \sigma_i). \tag{S7}$$

S3.3 Model specification

3.3.1 Bayesian

Derivation of d. The log posterior ratio d is equivalent to the log likelihood ratio plus an additive term representing the prior probability over category:

$$d = \log \frac{p(C=1 \mid x)}{p(C=2 \mid x)} = \log \frac{p(x \mid C=1)}{p(x \mid C=2)} + \log \frac{p(C=1)}{p(C=2)}. \tag{S8}$$

To get d, we need to find the expressions for the orientation measurement likelihood $p(x \mid C)$. The observer knows that the measurement x is caused by the stimulus s, but has no knowledge of s. Therefore, the optimal observer marginalizes over s:

$$p(x \mid C) = \int p(x \mid s) p(s \mid C)\, \mathrm{d}s. \tag{S9}$$

We substitute the expressions for the noise distribution and the stimulus distribution, and evaluate the integral:

$$p(x \mid C) = \int \mathcal{N}(s; x, \sigma^2) \mathcal{N}(s; \mu_C, \sigma_C^2)\, \mathrm{d}s = \mathcal{N}(x; \mu_C, \sigma^2 + \sigma_C^2). \tag{S10}$$

Plugging in the category-specific μ_C and σ_C, and substituting these expressions back into Equation S8, we get:

$$d = \frac{1}{2} \log \frac{\sigma^2 + \sigma_2^2}{\sigma^2 + \sigma_1^2} - \frac{\sigma_2^2 - \sigma_1^2}{2(\sigma^2 + \sigma_1^2)(\sigma^2 + \sigma_2^2)} x^2 + \log \frac{p(C=1)}{p(C=2)}. \tag{S11}$$

The 8 possible category and confidence responses are determined by comparing the log posterior ratio d to a set of decision boundaries $(k_0, k_1, \ldots, k_8)$. k_4 is equal to the observer's believed log prior ratio $\log \frac{p(C=1)}{p(C=2)}$, which functions as the boundary on d between the 4 category 1 responses and the 4 category 2 responses and is fit to capture possible category bias. k_4 is the only boundary parameter in models of category choice only (and not confidence). k_0 is fixed at $-\infty$ and k_8 is fixed at ∞. The observer chooses category 1 when d is positive. Thus there were 7 free boundary parameters: $(k_1, k_2, \ldots, k_7) = \mathbf{k}$.

The posterior probability of category 1 can be written as as $p(C=1 \mid x) = \frac{1}{1+\exp(-d)}$.

Decision boundaries. In the Bayesian models with d noise, we assume that, for each trial, there is an added Gaussian noise term on d, $\eta_d \sim p(\eta_d)$, where $p(\eta_d) = \mathcal{N}(0, \sigma_d^2)$, and σ_d is a free parameter. We pre-computed 101 evenly spaced draws of η_d and their corresponding probability densities $p(\eta_d)$. We used Equation S11 to compute a lookup table containing the values of d as a function of x, σ, and η_d. We then used linear interpolation to find sets of measurement boundaries $\mathbf{b}(\sigma)$ corresponding to each draw of η_d[17]. We then computed 101 response probabilities for each trial (as described in **Section S3.2**), one for each draw of η_d, and computed the weighted average according to $p(\eta_d)$. This gave the values of $p_{m,\theta}(r_i \mid s_i, \sigma_i)$ for each trial i, which are needed in order to compute the total log likelihood of the dataset under the model.

In the Bayesian choice model without d noise, we translate the decision boundary k_4 from a log prior ratio to a measurement boundary corresponding to the fitted noise levels σ. To do this, we use k_4 as the left-hand side of Equation S11 and solve for x at the fitted levels of σ. We used this model only for the purpose of obtaining estimates of the category decision boundary parameters, and not for model comparison.

3.3.2 Fixed

In the Fixed model, the observer compares the measurement to a set of boundaries that are not dependent on σ. We fit free parameters $\mathbf{k}$ and use measurement boundaries $b_r = k_r$.

3.3.3 Linear and Quadratic

In the Linear and Quadratic models, the observer compares the measurement to a set of boundaries that are linear or quadratic functions of σ. We fit free parameters $\mathbf{k}$ and $\mathbf{m}$ and use measurement boundaries $b_r(\sigma) = k_r + m_r\sigma$ (Linear) or $b_r(\sigma) = k_r + m_r\sigma^2$ (Quadratic).

3.3.4 Free

To estimate the category boundaries with minimal assumptions, we fit a Free model in which the observer compares the orientation measurement to a set of boundaries that vary nonparametrically (i.e., free of a parametric relationship with σ) across attention conditions. As with the Bayesian choice model without d noise (**Section S3.3.1**), we used this model only for the purpose of obtaining estimates of the category decision boundary parameters and did not fit confidence. We fit free parameters $k_{4,\text{valid}}$, $k_{4,\text{neutral}}$, $k_{4,\text{invalid}}$, and used measurement boundaries $b_{4,\text{attention condition}} = k_{4,\text{attention condition}}$.

S3.4 Lapse rates

In category and confidence models, we fit three different types of lapse rate. On each trial, there is some fitted probability of:

- A "full lapse" in which the category report is random, and confidence report is chosen from a distribution over the four levels defined by λ_1, the probability of a "very low confidence" response, and λ_4, the probability of a "very high confidence" response, with linear interpolation for the two intermediate levels.
- A "confidence lapse" $\lambda_{\text{confidence}}$ in which the category report is chosen normally, but the confidence report is chosen from a uniform distribution over the four levels.
- A "repeat lapse" λ_{repeat} in which the category and confidence response is simply repeated from the previous trial.

In category choice models, we fit a standard category lapse rate λ, as well the above "repeat lapse" λ_{repeat}.

S3.5 Parameterization

All parameters that defined the width of a distribution ($\sigma_{\text{valid}}, \sigma_{\text{neutral}}, \sigma_{\text{invalid}}, \sigma_d$) were sampled in log-space and exponentiated during the computation of the log likelihood. See **Table S1** for a complete list of model parameters for category choice and confidence models and **Table S3** for choice-only models.

S3.6 Model fitting

Rather than find a maximum likelihood estimate of the parameters, we sampled from the posterior distribution over parameters, $p(\theta \mid \text{data})$; this has the advantage of maintaining a measure of uncertainty about the parameters, which can be used both for model comparison and for plotting model fits. To sample from the posterior, we use an expression for the log posterior

$$\log p(\theta \mid \text{data}) = \log p(\text{data} \mid \theta) + \log p(\theta) + \text{constant}, \tag{S12}$$

where $\log p(\text{data} \mid \theta)$ is given in Equation S7. We assumed a factorized prior over each parameter j:

$$\log p(\theta) = \sum_{j=1}^{n} \log p(\theta_j), \tag{S13}$$

where j is the parameter index and n is the number of parameters. We took uniform (or, for parameters that were standard deviations, log-uniform) priors over reasonable, sufficiently large ranges[17], which we chose before fitting any models.

We sampled from the probability distribution using a Markov Chain Monte Carlo (MCMC) method, slice sampling[18]. For each model and dataset combination, we ran between 4 and 10 parallel chains with random starting points. For each chain, we took 100,000 to 1,000,000 total samples (depending on model computational time) from the posterior distribution over parameters. We discarded the first third of the samples and kept 6,667 of the remaining samples, evenly spaced to reduce autocorrelation. All samples with log posteriors more than 40 below the maximum log posterior were discarded. Marginal probability distributions of the sample log likelihoods were visually checked for convergence across chains. In total we had 120 model and dataset combinations, with a median of 40,002 kept samples (interquartile range = 13,334).

S3.7 Model comparison

3.7.1 Metric choice

To compare model fits while accounting for the complexity of each model, we computed an approximation of leave-one-out cross-validation. Leave-one-out cross-validation is the most thorough way to cross-validate but is very computationally intensive; it requires fitting the model t times, where t is the number of trials. The Pareto smoothed importance sampling approximation of leave-one-out cross-validation (PSIS-LOO, referred to here simply as LOO) takes into account the model's uncertainty landscape by using samples from the full posterior of θ[19].

LOO is currently the most accurate approximation of leave-one-out cross-validation[20].

We determined that our results were not dependent on our choice of model comparison metric. We computed AIC, BIC, AICc, WAIC[21], and LOO for all models in the 2 model groupings (category choice-plus-confidence and category choice-only), multiplying the non-LOO metrics by $-\frac{1}{2}$ to match the scale of LOO. For AIC, BIC, and AICc, we selected the MCMC sample with the highest log likelihood as our maximum-likelihood parameter estimate. Then we computed Spearman's rank correlation coefficient for every possible pairwise comparison of model comparison metrics for all model and dataset combinations, producing 20 total values (2 model groupings × 10 possible pairwise comparisons of model comparison metrics). All values were greater than 0.998, indicating that, had we used an information criterion instead of LOO, we would not have changed our conclusions. Furthermore, there are no model groupings in which the identities of the lowest- and highest-ranked models are dependent on the choice of metric. The agreement of these metrics strengthens our confidence in our conclusions.

3.7.2 Metric aggregation

In all figures where we present model comparison results (**Figures 3d, S3c, S5b**), we aggregate LOO scores by the following procedure: Choose a reference model (e.g. Fixed). Subtract all LOO scores from the corresponding observer's score for that model; this converts all scores to a LOO "difference from reference" score, with lower (more negative) indicating a better score and higher (more positive) indicating a worse score. Repeat the following standard bootstrap procedure 10,000 times: Choose randomly, with replacement, a group of datasets equal to the total number of unique datasets, and take the mean of their "difference from reference" scores for each model. Blue lines and shaded regions in model comparison plots indicate the median and 95% CI on the distribution of these bootstrapped mean "difference from reference" scores.

S3.8 Visualization of model fits

Model fits were plotted by bootstrapping synthetic group datasets with the following procedure: For each model and observer, we generated 20 synthetic datasets, each using a different set of parameters sampled, without replacement, from the posterior distribution of parameters. Each synthetic dataset was generated using the same stimuli as the ones presented to the real observer. We randomly selected a number of synthetic datasets equal to the number of observers to create a synthetic group dataset. For each synthetic group dataset, we computed the mean response per orientation bin. We then repeated this 1,000 times and computed the mean and standard deviation of the mean output per bin across all 1,000 synthetic group datasets, which we then plotted as the shaded regions. Therefore, shaded regions represent the mean ± 1 SEM of synthetic group datasets.

For plots with stimulus orientation on the horizontal axis (**Figures 2b, 3c, S3b, S5a**), orientation was binned according to quantiles of the stimulus distributions so that each point consisted of roughly the same number of trials. We took the overall stimulus distribution $p(s) = \frac{1}{2}\left(p(s \mid C=1) + p(s \mid C=2)\right)$ and found bin edges such that the probability mass of $p(s)$ was the same in each bin. We then plotted the binned data with linear spacing on the horizontal axis.

S3.9 Model recovery analysis

We performed a model recovery analysis[22] to test our ability to distinguish our choice and confidence models. We generated synthetic datasets from each model, using the same sets of stimuli that were originally randomly generated for each of the 12 observers. To ensure that the statistics of the generated responses were similar to those of the observers, we generated responses to these stimuli from 8 of the randomly chosen parameter estimates obtained via MCMC sampling (as described in **Section S3.6**) for each observer and model. In total, we generated 384 datasets (4 generating models × 12 observers × 8 datasets). We then fit all four models to every dataset, using maximum likelihood estimation (MLE) of parameters by an interior-point constrained optimization (MATLAB's *fmincon*), and computed AIC scores from the resulting fits. For reasons of computational tractability, we used AIC instead of LOO as the model comparison metric. Because AIC and LOO scores gave us near-identical model rankings for data from real subjects (**Section S3.7.1**), we do not believe that the model recovery results are dependent on choice of metric.

We found that the true generating model was the best-fitting model, on average, in all cases (**Figure S4**). Overall, AIC "selected" the correct model (i.e., AIC scores were lowest for the model that generated the data) for 87.5% of the datasets, indicating that our models are distinguishable.

Supplementary Figures

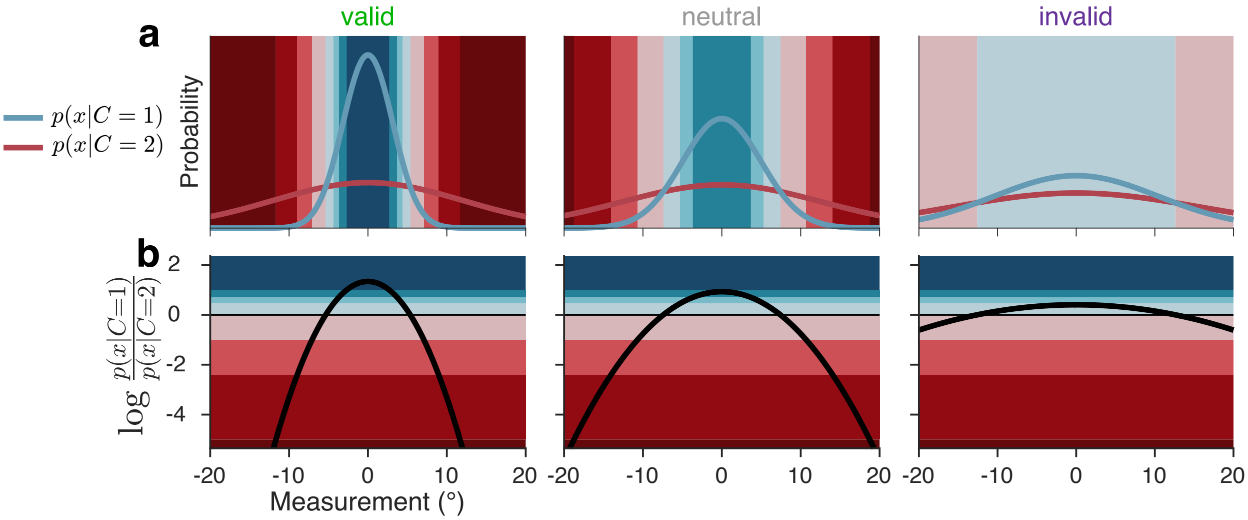

Figure S2: The Bayesian mapping from orientation measurement and attention-dependent uncertainty to response. Colors correspond to category and confidence response as in **Figure 1b**. (**a**) Blue and red curves show likelihood functions for the category distributions under example levels of uncertainty. (**b**) The Bayesian model maps measurement and uncertainty onto the decision variable, the log likelihood ratio (black curve). When the relative likelihood of category 1 is high, the decision variable is large and positive; when the relative likelihood of category 2 is high, it is large and negative. Response is determined by comparing the decision variable to boundaries that are fixed in log-likelihood-ratio space, but in measurement space vary as a function of uncertainty.

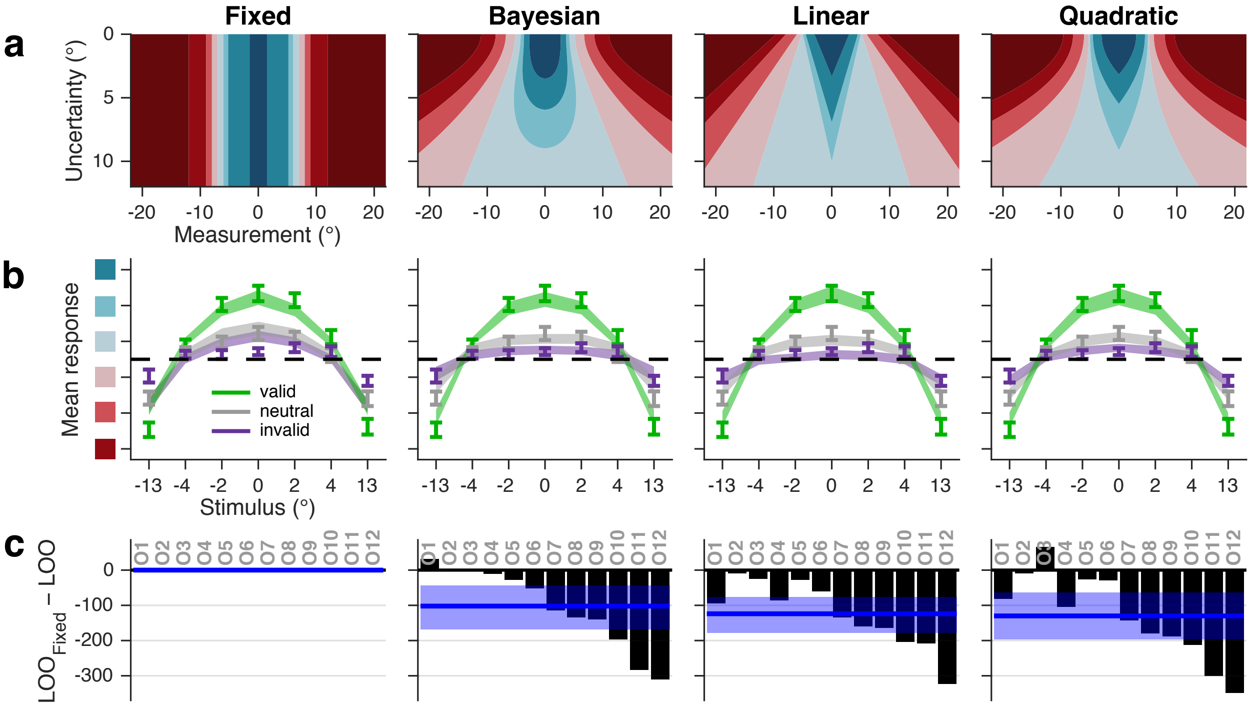

Figure S3: Category and confidence models. (**a**) Theoretical relation between orientation uncertainty and category and confidence decision boundaries for all models. (**b**) Mean response as a function of orientation and cue validity, as in **Figure 3c**. Stimulus orientation is binned to approximately equate the number of trials per bin. (**c**) Model comparison. Black bars represent individual observer LOO score differences of each model from Fixed. Negative values indicate that the corresponding model had a higher (better) LOO score than Fixed. Blue line and shaded region show median and 95% confidence interval of bootstrapped mean LOO differences across observers.

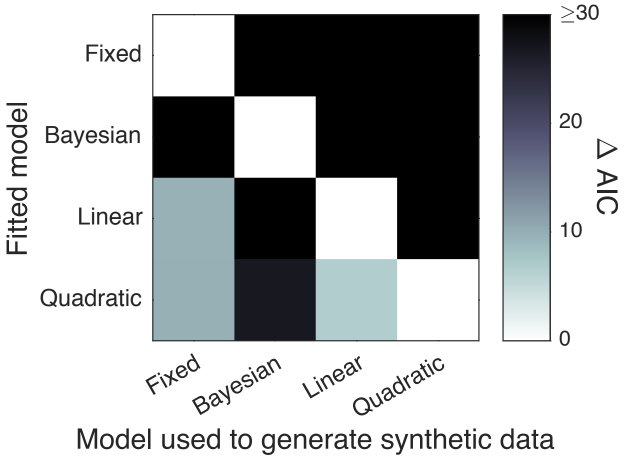

Figure S4: Model recovery analysis. Shade represents the difference between the mean AIC score (across synthetic datasets) for each fitted model and for the one with the lowest mean AIC score. White squares indicate the model that had the lowest mean AIC score when fitted to data generated from each model. The fact that all white squares lie on the diagonal indicates that the true generating model was the best-fitting model, on average, in all cases.

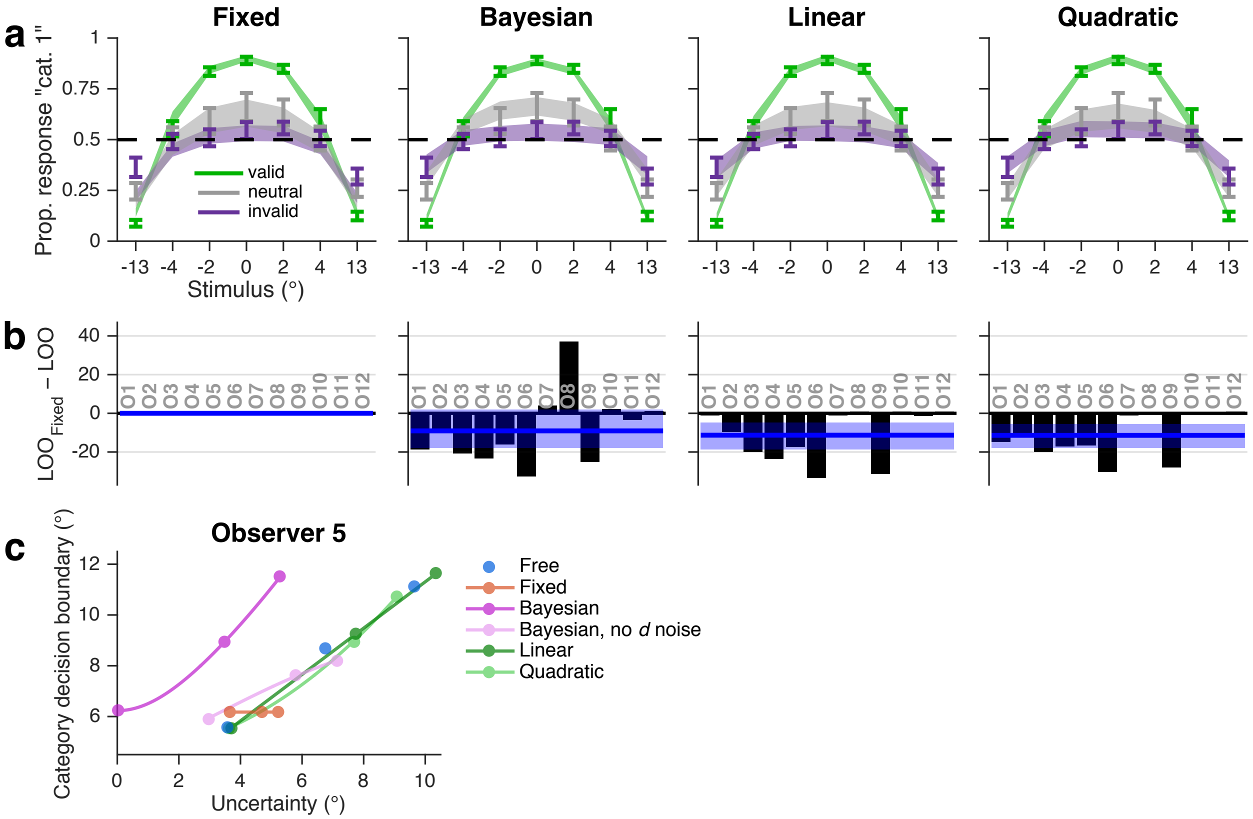

Figure S5: Category choice-only models. (**a**) Proportion of category 1 responses as a function of orientation and cue validity. Error bars show mean and SEM across observers. Shaded regions are mean and SEM of model fits (**Section S3.8**). Stimulus orientation is binned to approximately equate the number of trials per bin. (**b**) LOO model comparison, as in **Figure S3c**. (**c**) Mean MCMC orientation uncertainty and category choice boundary parameter estimates for a representative observer. Estimates are plotted as a function of attention condition (valid, neutral, invalid; filled circles), along with their generating functions (curves), for the four main models fit to the category choice data only, plus a Bayesian model with no noise on the decision variable d and a nonparametric model in which choice boundaries are unconstrained (Free; parameter estimates from this model are plotted in gray for all subjects in **Figure 4**). The Bayesian curve is to the left of the other curves, because noise attributed to orientation uncertainty in the other models is partially attributed to decision noise in the Bayesian model; when the decision noise parameter is removed (Bayesian, no d noise), the curve aligns with the others.

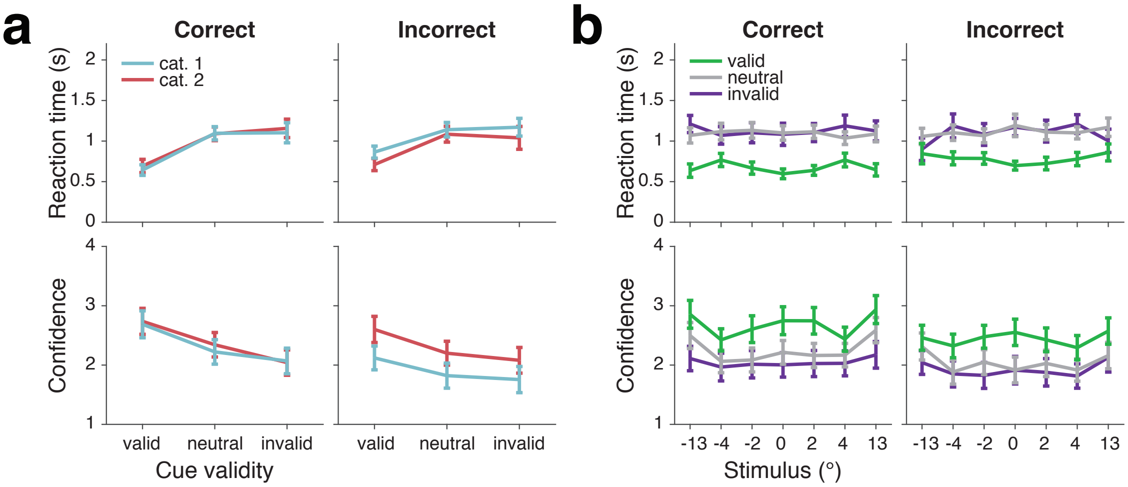

Figure S6: RT and confidence data broken down by category and accuracy. RT did not depend strongly on category or accuracy, though it was slightly longer for valid incorrect compared to valid correct trials. Confidence was higher overall for correct compared to incorrect trials. Confidence was higher and RT slightly faster for category 2 incorrect trials compared to category 1 incorrect trials, likely because there are more category 2 trials with high probability of being category 1 (which would lead to a high confidence error) than category 1 trials with high probability of being category 2.

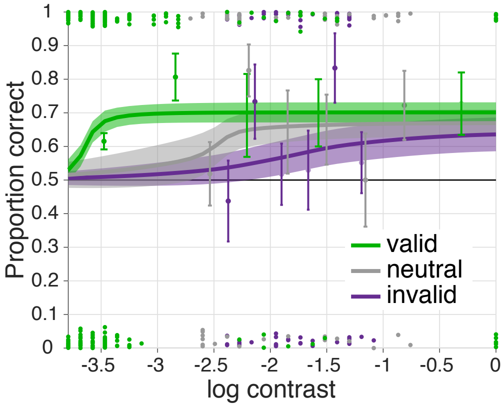

Figure S7: Example plot used to determine per-observer stimulus contrast. Each curve shows the mean ±1 SD of the posterior over psychometric functions for each attention condition. Error bars indicate the mean ±1 SD of the beta distribution over correctness within log contrast bins. A dot indicates one correct or incorrect trial, located respectively at the top or bottom of the plot, with vertical jitter. For this example observer, we selected a natural log contrast of -2.3 (i.e., a contrast of 10%).

Supplementary Tables

	Fixed	Bayesian	Linear	Quadratic
Measurement noise	$\sigma_{\text{valid}}, \sigma_{\text{neutral}}, \sigma_{\text{invalid}}$			
Orientation-dependent noise	ψ			
Decision boundaries	k_{1-7}		k_{1-7}, m_{1-7}	
d noise		σ_d		
Lapse rates	λ_1, λ_4, $\lambda_{\text{confidence}}$, λ_{repeat}			
Total number of parameters	15	16	22	22

Table S1: Parameters of category choice and confidence decision models.

		15 pars.	16 pars.	22 pars.
		Fixed	Bayesian	Linear
22 pars.	Quadratic	129 $[65, 198]$	27 $[0, 53]$	5 $[-18, 28]$
22 pars.	Linear	124 $[77, 177]$	21 $[-3, 48]$	
16 pars.	Bayesian	102 $[45, 167]$		

Table S2: Cross comparison of all category choice and confidence decision models. Cells indicate medians and 95% CI of bootstrapped mean LOO score differences. A positive median indicates that the model in the corresponding row had a higher score (better fit) than the model in the corresponding column.

	Fixed	Bayesian	Bayesian, no d noise*	Linear	Quadratic	Free*
Measurement noise	$\sigma_{\text{valid}}, \sigma_{\text{neutral}}, \sigma_{\text{invalid}}$					
Orientation-dependent noise	ψ					
Decision boundaries	k			k, m		$k_{\text{valid}}, k_{\text{neutral}}, k_{\text{invalid}}$
d noise		σ_d				
Lapse rates	λ, λ_{repeat}					
Total number of parameters	7	8	7	8	8	9

Table S3: Parameters of category choice-only decision models. * indicates models that were used only for obtaining parameter estimates (**Figures 4, S5c**), and not for model comparison.

		7 pars.	8 pars.	8 pars.
		Fixed	Bayesian	Linear
8 pars.	Quadratic	11 $[5, 18]$	2 $[-2, 9]$	0 $[-2, 3]$
8 pars.	Linear	11 $[4, 19]$	2 $[-3, 10]$	
8 pars.	Bayesian	9 $[-2, 18]$		

Table S4: Cross comparison of all category choice-only decision models. Conventions as in **Table S2**.

References

[1] Qamar, A. T. *et al.* Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proc Nat Acad Sci USA* **110**, 20332–20337 (2013).

[2] Green, D. M. & Swets, J. A. *Signal detection theory and psychophysics* (Wiley, New York, 1966).

[3] Gorea, A. & Sagi, D. Failure to handle more than one internal representation in visual detection tasks. *Proc Natl Acad Sci USA* **97**, 12380–12384 (2000).

[4] Gorea, A. & Sagi, D. Disentangling signal from noise in visual contrast discrimination. *Nat Neurosci* **4**, 1146–1150 (2001).

[5] Kontsevich, L. L., Chen, C.-C., Verghese, P. & Tyler, C. W. The unique criterion constraint: a false alarm? *Nat Neurosci* **5**, 707 (2002).

[6] Lu, Z. L. & Dosher, B. A. External noise distinguishes attention mechanisms. *Vis Res* **38**, 1183–1198 (1998).

[7] Carrasco, M., Penpeci-Talgar, C. & Eckstein, M. Spatial covert attention increases contrast sensitivity across the CSF: support for signal enhancement. *Vis Res* **40**, 1203–1215 (2000).

[8] Dosher, B. A. & Lu, Z. L. Noise exclusion in spatial attention. *Psych Sci* **11**, 139–146 (2000).

[9] Adler, W. T. & Ma, W. J. Comparing Bayesian and non-Bayesian accounts of human confidence reports. *bioRxiv* 093203 (2018). related:qMQe6XVts8AJ.

[10] Aitchison, L., Bang, D., Bahrami, B. & Latham, P. E. Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Comput Biol* **11**, e1004519 (2015).

[11] Pelli, D. G. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat Vis* **10**, 437–442 (1997).

[12] Brainard, D. H. The Psychophysics Toolbox. *Spat Vis* **10**, 433–436 (1997).

[13] Kleiner, M., Brainard, D. H. & Pelli, D. G. What's new in Psychtoolbox-3? ECVP Abstract Supplement . *Perception* **36** (2007).

[14] Kontsevich, L. L. & Tyler, C. W. Bayesian adaptive estimation of psychometric slope and threshold. *Vis Res* **39**, 2729–2737 (1999).

[15] Prins, N. The psychometric function: the lapse rate revisited. *JOV* **12**, 25–25 (2012).

[16] Girshick, A. R., Landy, M. S. & Simoncelli, E. P. Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nat Neurosci* **14**, 926–932 (2011).

[17] Acerbi, L., Vijayakumar, S. & Wolpert, D. M. On the origins of suboptimality in human probabilistic inference. *PLoS Comput Biol* **10**, e1003661 (2014).

[18] Neal, R. M. Slice sampling. *Ann Stat* **31**, 705–741 (2003).

[19] Vehtari, A., Gelman, A. & Gabry, J. Efficient implementation of leave-one-out cross-validation and WAIC for evaluating fitted Bayesian models. *arXiv* 1507.04544v1 (2015). 3706900867636205788related:3BgL-bKOcTMJ.

[20] Acerbi, L., Dokka, K., Angelaki, D. E. & Ma, W. J. Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *bioRxiv* e150052 (2017).

[21] Gelman, A., Hwang, J. & Vehtari, A. Understanding predictive information criteria for Bayesian models. *Stat Comput* **24**, 997–1016 (2014).

[22] van den Berg, R., Awh, E. & Ma, W. J. Factorial comparison of working memory models. *Psych Rev* **121**, 124–149 (2014).