Humans flexibly 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

Perceptual decisions are better when they take uncertainty into account. Here we show that human decision-making accounts for uncertainty arising not only from external factors, but from the observer's cognitive state. We manipulated uncertainty in an orientation categorization task from trial to trial using only an attentional cue. Category and confidence decision boundaries shifted in an approximately Bayesian fashion. This responsiveness likely improves perceptual decisions in natural vision.

Sensory representations are inherently noisy. In vision, stimulus factors such as low contrast, blur, and visual noise can increase an observer's uncertainty about a visual stimulus. Optimal perceptual decision-making requires taking into account both the sensory measurements and their associated uncertainty [1]. When driving on a foggy day, for example, you may be more uncertain about the distance between your car and the car in front of you than you would be on a clear day, and try to keep further back. Humans often respond to sensory uncertainty in this way [2,3], adjusting both their choice [4] and confidence [5] behavior. Confidence is a metacognitive measure that reflects the observer's degree of certainty about a perceptual decision [6,7].

Uncertainty, however, often arises not from the external world but from one's internal state. Attention is a key internal state variable that governs the uncertainty of visual representations [8,9]; it modulates basic perceptual properties like contrast sensitivity [10,11] and spatial resolution [12]. Interactions between attention and perceptual decision-making have been of interest [13–19], but the formal modeling tools used to investigate uncertainty in perception have not yet been applied to the uncertainty due to one's attentional state. It is therefore unknown whether and how humans take attention-dependent uncertainty into account when making perceptual decisions. Here we combined psychophysical experiments with modeling of the experimental data and model comparison to determine the influence of attention on perceptual decision rules.

Observers categorized drifting grating stimuli as drawn from either a narrow distribution around horizontal (SD = 3°, category 1) or a wide distribution around horizontal (SD = 12°, category 2) (**Figure 1a**) [4]. This task requires distinguishing a more specific from a more general perceptual category, similar to determining whether an approaching person is someone you know. Four stimuli were briefly presented on each trial, and a response cue indicated which stimulus to report. Observers reported both their category choice (category 1 vs. 2) and their degree of confidence on a 4-point scale using a single button press (**Figure 1b**). Using a single button press prevented post-choice influences on the confidence judgment [20] and emphasized that confidence should reflect the observer's perception rather than a preceding motor response. Twelve observers participated, with about 2000 trials per observer.

We manipulated voluntary (endogenous) attention on a trial-to-trial basis using a spatial cue that pointed to either one stimulus location (valid condition when the response cue matched the cue, 66.7% of trials; and invalid condition when it did not match, 16.7% of trials) or all four locations (neutral condition, 16.7% of

[*] Equal author contribution

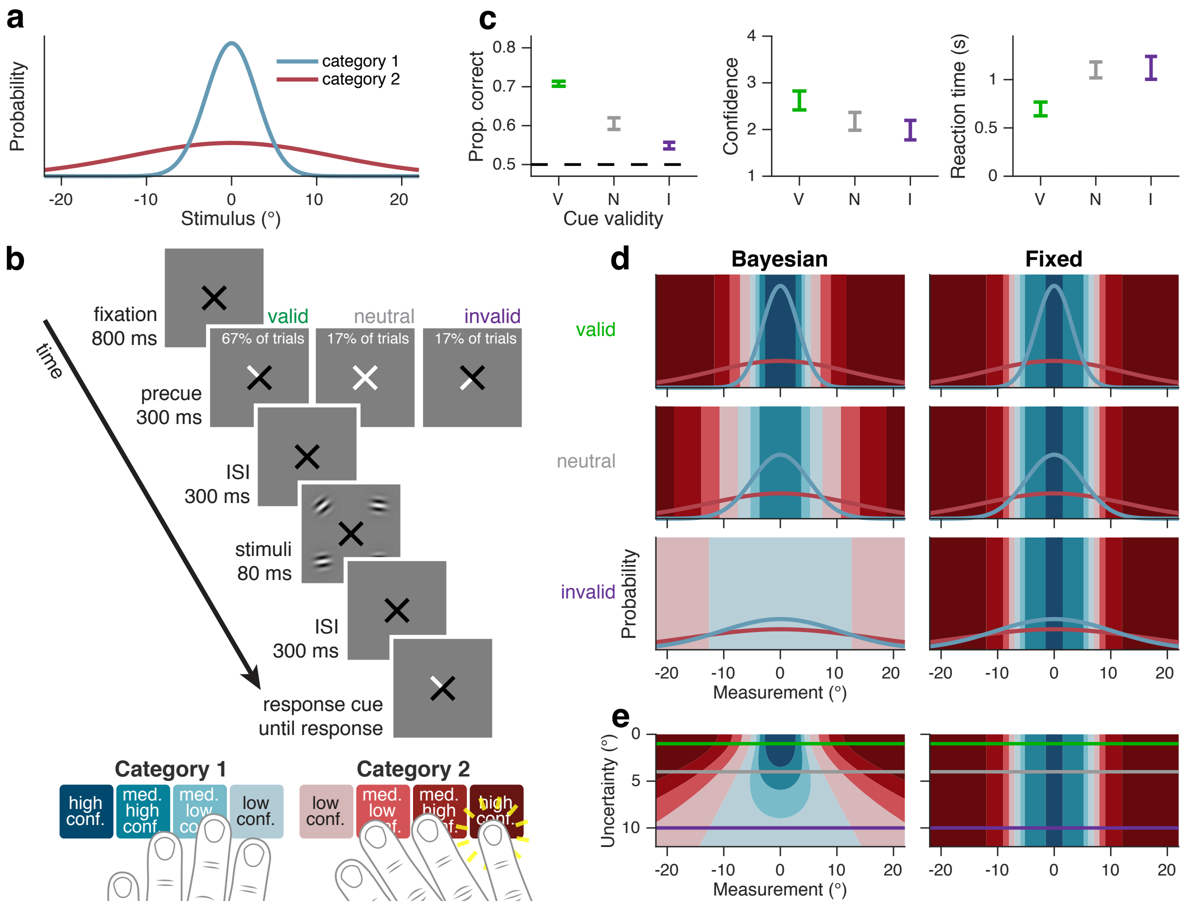

Figure 1: Stimuli and task. (**a**) Stimulus orientation distributions for each category. (**b**) Trial sequence. Cue validity, the likelihood that a precue to one quadrant would match the response cue, was 80%. (**c**) Behavioral performance, trial-weighted mean and SEM across observers for valid, neutral, and invalid precue conditions. Maximum accuracy is ~80% because the stimulus distributions overlap. $n = 12$ observers. (**d**) Schematic of Bayesian (left) and Fixed (right) models. As attention decreases, uncertainty (the measurement noise SD) increases, and orientation measurement likelihoods (blue and red curves) widen[21]. In the Bayesian model, choice and confidence boundaries are defined by posterior probability ratios and therefore change as a specific function of uncertainty. In the Fixed model, boundaries do not depend on uncertainty. (**e**) Decision rules for Bayesian and Fixed models show the mappings from orientation measurement and uncertainty to category and confidence responses (color code in **b**). Horizontal lines indicate the uncertainty levels used in **d**; note that the regions intersecting with a horizontal line match the regions in the corresponding plot in **d**.

trials) (**Figure 1b**). Cue validity increased categorization accuracy [one-way repeated-measures ANOVA, $F(2, 11) = 95.88$, $p < 10^{-10}$], with higher accuracy following valid cues [two-tailed paired t-test, $t(11) = 7.92$, $p < 10^{-5}$] and lower accuracy following invalid cues [$t(11) = 4.62$, $p < 10^{-3}$], relative to neutral cues (**Figure 1c**). This pattern confirms that attention increased orientation sensitivity (e.g.,[11,22]). Attention also increased confidence ratings [$F(2, 11) = 13.35$, $p < 10^{-3}$] and decreased reaction time [$F(2, 11) = 28.76$, $p < 10^{-6}$], ruling out speed-accuracy tradeoffs as underlying the effect of attention on accuracy (**Figure 1c**).

We assessed whether observers changed their category and confidence decision boundaries to account for attention-dependent orientation uncertainty by fitting two main models. In the Bayesian model, observers adjust their decision boundaries in orientation space as uncertainty changes[4,5]. Decisions depend on the relative posterior probabilities of the two categories (**Figures 1d,e, S1**), a strategy that maximizes accuracy and produces confidence reports that are a function of the posterior probability of being correct. In the Fixed model, observers use the same decision criteria, regardless of the attended location (**Figure 1d,e**)[13–16,23–27].

The models are differentiated by how they map attention condition and stimulus orientation onto a response. We therefore plotted behavior as a function of these two variables. Performance was a "W"-shaped function of stimulus orientation (**Figure 2a**), reflecting the greater difficulty in categorizing a stimulus when its orientation was near a category boundary. Attention increased the sensitivity of category and confidence responses to the stimulus orientation (**Figure 2b**, error bars). We used Markov Chain Monte Carlo sampling to fit models to raw, trial-to-trial category and confidence responses from each observer (**Methods**, **Table S1**).

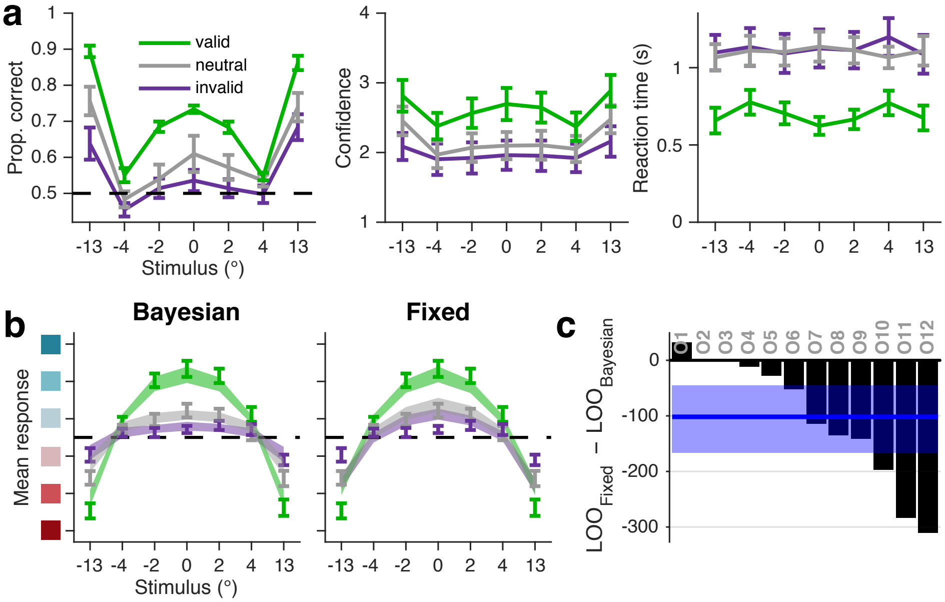

Figure 2: Behavioral data and model fits. (**a**) Accuracy, confidence ratings, and reaction time as a function of orientation and cue validity. Error bars show trial-weighted mean and SEM across observers. (**b**) Mean response as a function of orientation and cue validity. Mean response is an 8-point scale ranging from "high confidence" category 1 to "high confidence" category 2, with colors corresponding to those in **Figure 1b**; only the middle 6 responses are shown. Error bars show mean and SEM across observers. Shaded regions are mean and SEM of model fits (**Methods**). (**c**) Model comparison. Black bars represent individual observer LOO differences of Bayesian from Fixed. Negative values indicate that Bayesian had a higher (better) LOO score than Fixed. Blue line and shaded region show median and 95% confidence interval of bootstrapped mean differences across observers.

Observers' decisions accounted for attention-dependent uncertainty. The Bayesian model captured the data well (**Figure 2b**) and substantially outperformed the Fixed model (**Figure 2c**), which had systematic deviations from the data (though the fit depends on the full data set, note deviations near zero tilt and at large tilts in **Figure 2b**). Bayesian outperformed Fixed by PSIS-LOO (an approximation of leave-one-out cross-validated log likelihood[28], henceforth LOO) differences (median and 95% CI of bootstrapped mean differences across observers) of 102 [45, 167].

To determine whether Bayesian computations are necessary to produce the behavioral data, we tested two models with heuristic decision rules in which the decision boundaries vary as linear or quadratic functions of uncertainty, approximating the Bayesian boundaries (**Figure S2a**). The Linear and Quadratic models both outperformed the Fixed model (LOO differences of 124 [77, 177] and 129 [65, 198], respectively; **Figure S2b,c**). The best model, quantitatively, was Quadratic[4,5]. **Table S2** shows all pairwise comparisons of the models. Model recovery showed that our models were meaningfully distinguishable (**Figure S3**). Decision rules therefore changed with attention without requiring Bayesian computations.

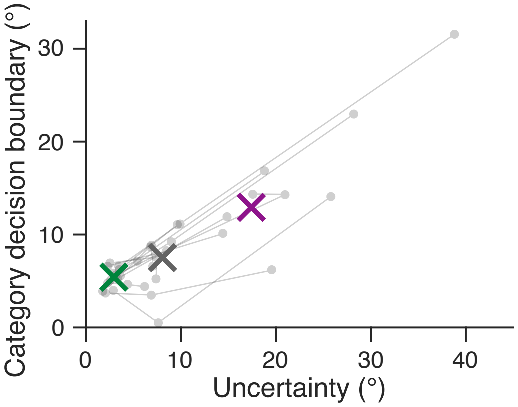

Figure 3: Free model analysis. Group mean MCMC parameter estimates (crosses) show systematic changes in the category decision boundary across attention conditions (green = valid, gray = neutral, purple = invalid). The same pattern can be seen for individual observers: each gray line corresponds to an observer, with connected gray points representing the estimates for valid, neutral, and invalid attention conditions, in that order. Note that each point represents a pair of parameter estimates: uncertainty and category decision boundary for a specific attention condition.

We next asked whether the category decision boundary alone—regardless of confidence—accounts for attention-dependent uncertainty. We were able to answer this question because, unlike in a traditional left vs. right orientation discrimination task, the optimal category decision boundaries in this task depend on orientation uncertainty (**Figure 1d,e**)[4]. We fit the four models to the category choice data only and again rejected the Fixed model (**Figure S4a,b**; **Tables S3, S4**). We also fit the category choice data with a Free model in which the category decision boundaries varied freely and independently for each attention condition. The estimated boundaries differed between valid and invalid trials (**Figures 3, S4c**), with a mean difference of 7.5° (SD = 7.8°) [two-tailed paired t-test, $t(11) = 3.33$, $p < 10^{-2}$]. Therefore, category criteria, independent of confidence criteria, varied as a function of attention-dependent uncertainty.

Our findings of flexible decision boundaries are surprising in light of the "unified criterion" account of perceptual decision-making[25,26]. According to this account, when multiple relevant items are simultaneously present in a display, the observer adopts a single, fixed decision boundary (the "unified criterion") that is used for all items, regardless of stimulus properties or attentional state. Findings in support of this account[13–16,23–27] suggest a rigid, suboptimal mechanism for perceptual decision-making in real-world scenarios, which inevitably involve multiple items. Previous studies may have been limited in their ability to distinguish changes in criteria from changes in internal signal variability[29]. Other differences between our study and those supporting a unified criterion include the type of decision required of the observer—categorization here versus detection and orthogonal discrimination* in previous studies[13,15,16,23–27]—and the metacognitive report of confidence here versus visibility used previously[13,30].

Few studies have examined the influence of attention on confidence, and findings have been mixed. Two studies found that attention increased confidence[31,32], but another found no effect of attention on confidence[33]. The latter result has been attributed to response speed pressures[31]. Two other studies suggested an inverse relation between attention and confidence: one reported lower confidence for cued than uncued error trials[19], and another found that lower fMRI BOLD activation in the dorsal attention network correlated with higher confidence[14]. Our results clearly support a positive relation between attention and confidence (**Figure 1c**) and that attention increases the sensitivity of confidence to fine-grained stimulus features (**Figure 2a**). Further, they reveal that this relation is approximately Bayesian and can arise from heuristic decision rules that are more straightforward to implement computationally than exact Bayesian inference.

The mechanisms for decision adjustment under attention-dependent uncertainty could be mediated by effective contrast[10,34,35]. Alternatively, attention-dependent decision-making may rely on higher-order monitoring of attentional state. Our finding that human observers incorporate attention-dependent uncertainty into perceptual decisions and confidence reports in a statistically appropriate fashion raises the question of what other kinds of cognitive or motivational states can be incorporated into perceptual decision-making. Attention is typically spread unevenly across multiple objects in a visual scene, so the ability to account for attention likely improves perceptual decisions in natural vision.

*Orthogonal discrimination is often used as a proxy for detection.

ACKNOWLEDGEMENTS

The authors would like to thank: Roshni Lulla and Gordon C. Bill for assistance with data collection and helpful discussions; Luigi Acerbi for helpful ideas and tools related to model fitting and model comparison; and Christopher Summerfield for comments on the manuscript. This material was based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1342536 (W.T.A.) and by the National Eye Institute of the National Institutes of Health under awards T32EY007136 and F32EY025533 (R.N.D.).

AUTHOR CONTRIBUTIONS

R.N.D, W.T.A., M.C., and W.J.M. designed the research and wrote the manuscript. W.T.A. and W.J.M. developed the models. R.N.D. and W.T.A. performed the experiments and analyzed data.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

DATA AND CODE AVAILABILITY

All data and code used for running experiments, model fitting, and plotting is available on GitHub (https://github.com/wtadler/confidence).

Methods

1 Experiment

1.1 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 experiment, either because they were excluded on the basis of their staircase performance (**Section 1.3.7**) or because they chose to stop participating before all sessions were completed. 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.

1.2 Apparatus and stimuli

1.2.1 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[36–38].

1.2.2 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.

1.2.3 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%.

1.2.4 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.

1.3 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 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%).

1.3.1 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.

1.3.2 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.

1.3.3 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.

1.3.4 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.

1.3.5 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.

1.3.6 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[39], extended to include the lapse rate[40]. 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.

1.3.7 Observer and contrast selection

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 observer was considered eligible for the remainder of the study if there existed a contrast for which the mean minus the SD of the posterior over invalid psychometric functions was above chance, and 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. 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 did not continue the study. We used this observer screening and contrast selection procedure because, in order to test our hypothesis, we needed uncertainty to depend on attention. This procedure increased the probability that uncertainty would vary between attention conditions in the final dataset. Selected contrasts ranged from 4% to 60% across observers.

1.3.8 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.

1.3.9 Confidence/attention testing

These were the main experimental blocks. The trial procedure was the same as in confidence/attention training blocks, but with no trial-to-trial feedback whatsoever. The trial sequence is shown in **Figure 1b**. 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.

2 Modeling

The modeling procedures were similar to those used by Adler and Ma[5]. 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.

2.1 Measurement noise

We used free parameters to characterize orientation measurement noise variance σ 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[41]. The measurement noise SD comes out to

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

2.2 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:

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

For this task, $b_0 = 0°$ and $b_8 = \infty°$. Since the task is symmetric around $0°$, we only use $|s|$ in our computation of the log likelihood.

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:

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

2.3 Model specification

2.3.1 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:

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

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:

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

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

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

Plugging in the category-specific μ_C and σ_C, and substituting these expressions back into equation (2), we get:

$$d = \frac{1}{2} \log \frac{\sigma^2 + \sigma_2^2}{\sigma^2 + \sigma_1^2} - \frac{\sigma_2^2 - \sigma_1^2}{2(\sigma^2 + \sigma_1^2)(\sigma^2 + \sigma_2^2)} x^2 + \log \frac{p(C=1)}{p(C=2)}. \tag{3}$$

The 8 possible category and confidence responses are determined by comparing the log posterior ratio d to a set of decision boundaries $\mathbf{k} = (k_0, k_1, \ldots, k_8)$. k_4 is equal to the 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; 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: $k_1, k_2, k_3, k_4, k_5, k_6, k_7$.

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 (3) 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[42]. We then computed 101 response probabilities for each trial (as described in **Section 2.2**), 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 (3) 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.

2.3.2 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 $b_r = k_r$.

2.3.3 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).

2.3.4 Free

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 2.3.1**), we used this model only for the purpose of obtaining estimates of the category decision boundary parameters. We fit free parameters $k_{4,\text{valid}}$, $k_{4,\text{neutral}}$, $k_{4,\text{invalid}}$, and used measurement boundaries $b_{4,\text{attention condition}} = k_{4,\text{attention condition}}$.

2.4 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}.

2.5 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.

2.6 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 unnormalized log posterior

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

where $\log p(\text{data} \mid \theta)$ is given in equation (1). We assumed a factorized prior over each parameter j:

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

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[43], which we chose before fitting any models.

We sampled from the probability distribution using a Markov Chain Monte Carlo (MCMC) method, slice sampling[44]. 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, discarded the first third of the samples, and kept 6,667 of the remaining samples, randomly selected. 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).

2.7 Model comparison

2.7.1 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[28]:

$$\text{LOO} = \sum_{i=1}^{t} \log \frac{\sum_u w_{i,u} p(r_i \mid \theta_u)}{\sum_u w_{i,u}},$$

where θ_u is the u-th sampled set of parameters, and $w_{i,u}$ is the importance weight of trial i for sample u. Pareto smoothed importance sampling provides an accurate and reliable estimate of the weights.[45] LOO is currently the most accurate approximation of leave-one-out cross-validation[46].

We determined that our results were not dependent on our choice of model comparison metric. We computed AIC, BIC, AICc, WAIC[47], 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.

2.7.2 Metric aggregation

In all figures where we present model comparison results (**Figures 2c, S2c, S4b**), 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.

2.8 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 2a,b, S2b, S4a**), 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.

2.9 Model recovery

We performed a model recovery analysis[48] 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 2.6**) 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 2.7.1**), 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 S3**). 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

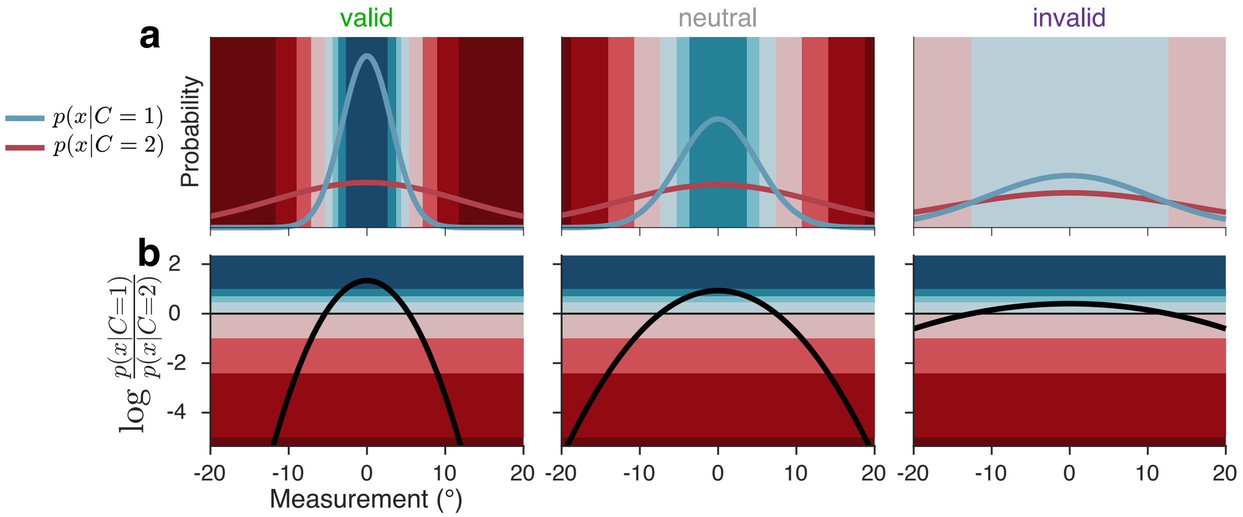

Figure S1: The Bayesian mapping from orientation measurement and attention-dependent uncertainty to response. Colors correspond to category and confidence response as in **Figure 1b**. (**a**) Blue and red curves show likelihood functions for the category distributions under example levels of uncertainty. (**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.

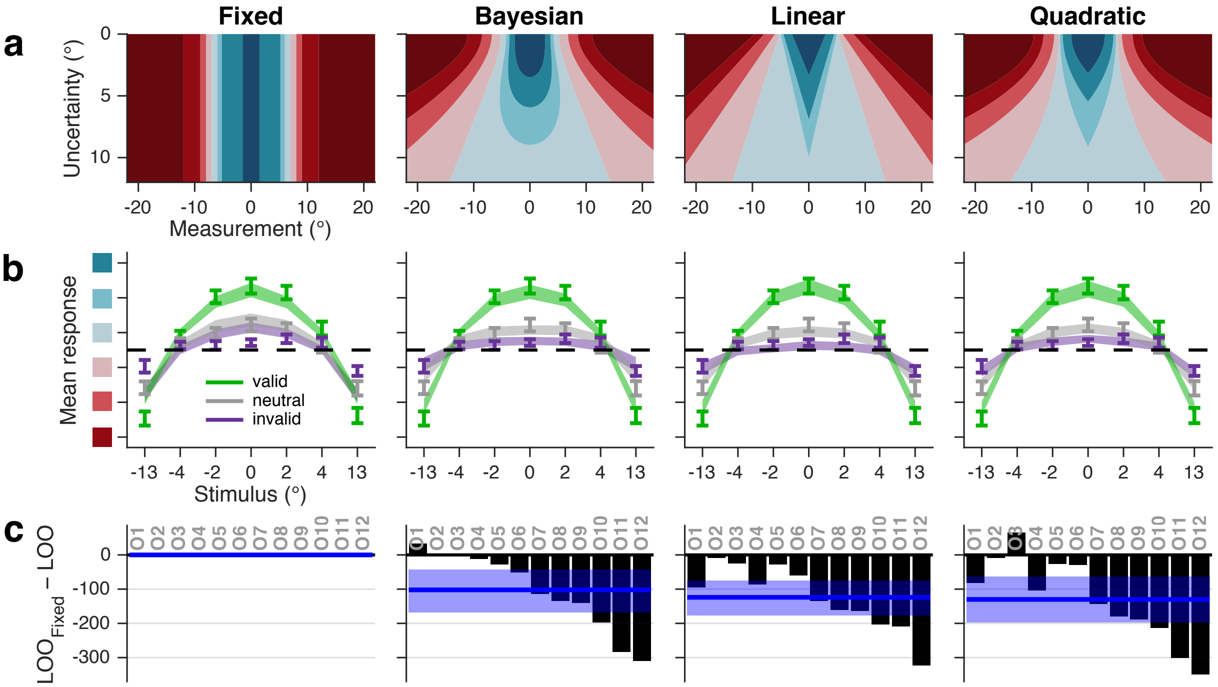

Figure S2: Category and confidence models. (**a**) Theoretical relation between orientation uncertainty and category and confidence decision boundaries for all models. (**b**) Mean response as a function of orientation and cue validity, as in **Figure 2b**. (**c**) Model comparison. Black bars represent 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.

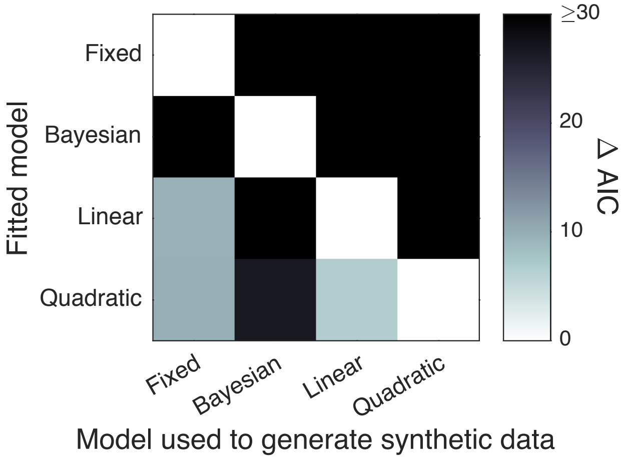

Figure S3: Model recovery analysis. Shade represents the difference between the mean AIC score (across synthetic datasets) for each fitted model and for the one with the lowest mean AIC score. White squares indicate the model that had the lowest mean AIC score when fitted to data generated from each 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.

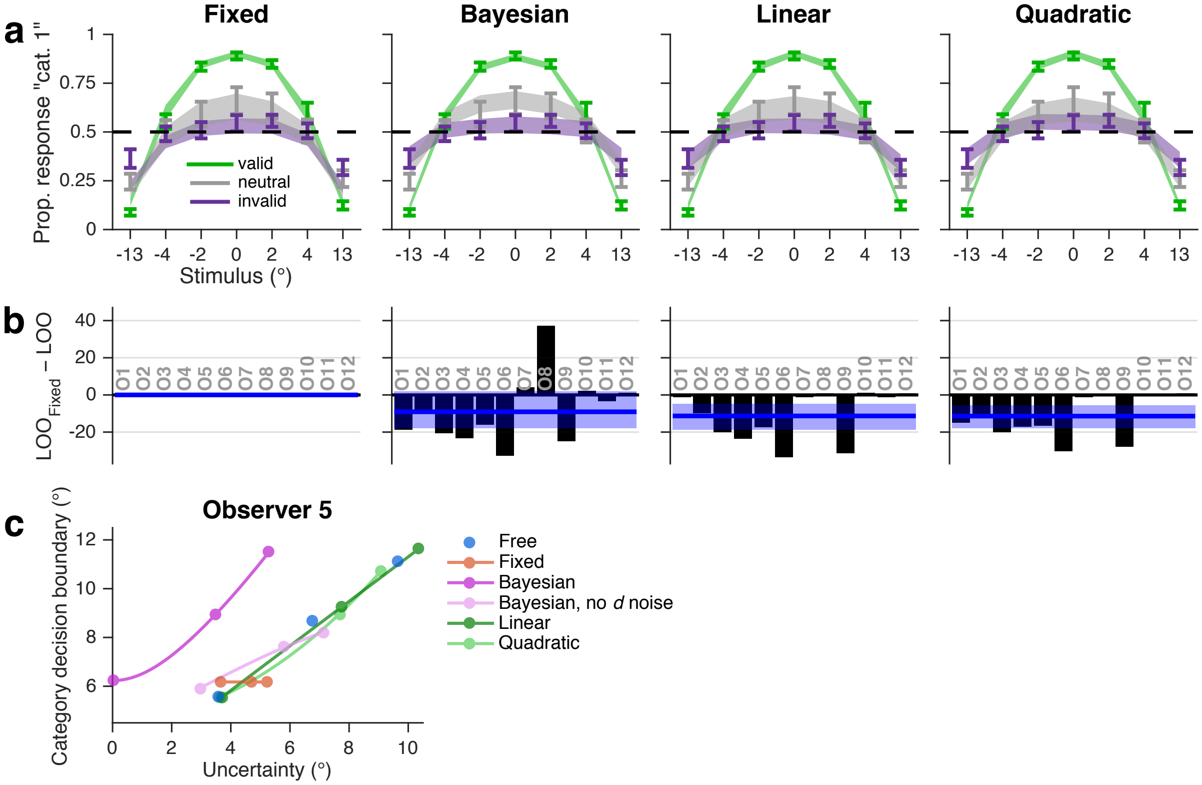

Figure S4: Category choice-only models. (**a**) Proportion of category 1 responses as a function of orientation and cue validity. Error bars show mean and SEM across observers. Shaded regions are mean and SEM of model fits (**Methods**). (**b**) LOO model comparison, as in **Figure S2c**. (**c**) Mean 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 3**). 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.

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	$\lambda_1, \lambda_4, \lambda_{\text{confidence}}, \lambda_{\text{repeat}}$			
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	$\lambda, \lambda_{\text{repeat}}$					
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 3, S4c**), 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] Knill, David C & Richards, W. *Perception as Bayesian Inference* (Cambridge University Press, Cambridge, UK, 1996).

[2] Trommershäuser, J., Kording, K. & Landy, M. S. (eds.) *Sensory Cue Integration* (Oxford University Press, Oxford, UK, 2011).

[3] Ma, W. J. & Jazayeri, M. Neural coding of uncertainty and probability. *Annu. Rev. Neurosci.* **37**, 205–220 (2014).

[4] Qamar, A. T. *et al.* Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proceedings of the National Academy of Sciences* **110**, 20332–20337 (2013).

[5] Adler, W. T. & Ma, W. J. Human confidence reports account for sensory uncertainty but in a non-Bayesian way. *bioRxiv* 093203 (2017). related:qMQe6XVts8AJ.

[6] Mamassian, P. Visual Confidence. *Annu. Rev. Vis. Sci.* **2**, 459–481 (2016).

[7] Fleming, S. M. & Daw, N. D. Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review* **124**, 91–114 (2017).

[8] Carrasco, M. Visual attention: the past 25 years. *Vision Research* **51**, 1484–1525 (2011).

[9] Reynolds, J. H. & Chelazzi, L. Attentional modulation of visual processing. *Annu. Rev. Neurosci.* **27**, 611–647 (2004).

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

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

[12] Anton-Erxleben, K. & Carrasco, M. Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence. *Nat Rev Neurosci* **14**, 188–200 (2013).

[13] Rahnev, D. *et al.* Attention induces conservative subjective biases in visual perception. *Nature Neuroscience* **14**, 1513–1515 (2011).

[14] Rahnev, D. A., Maniscalco, B., Luber, B., Lau, H. & Lisanby, S. H. Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *J Neurophysiol* **107**, 1556–1563 (2012).

[15] Morales, J. *et al.* Low attention impairs optimal incorporation of prior knowledge in perceptual decisions. *Atten Percept Psychophys* **77**, 2021–2036 (2015).

[16] Caetta, F. & Gorea, A. Upshifted decision criteria in attentional blink and repetition blindness. *Visual Cognition* **18**, 413–433 (2010).

[17] Schoenherr, J. R., Leth-Steensen, C. & Petrusic, W. M. Selective attention and subjective confidence calibration. *Atten Percept Psychophys* **72**, 353–368 (2010).

[18] Sherman, M. T., Seth, A. K., Barrett, A. B. & Kanai, R. Prior expectations facilitate metacognition for perceptual decision. *Consciousness and Cognition* **35**, 53–65 (2015).

[19] Baldassi, S., Megna, N. & Burr, D. C. Visual clutter causes high-magnitude errors. *Plos Biol* **4**, e56 (2006).

[20] Navajas, J., Bahrami, B. & Latham, P. E. Post-decisional accounts of biases in confidence. *Current Opinion in Behavioral Sciences* **11**, 55–60 (2016).

[21] Giordano, A. M., McElree, B. & Carrasco, M. On the automaticity and flexibility of covert attention: a speed-accuracy trade-off analysis. *JOV* **9**, 30.1–10 (2009).

[22] Cameron, E. L., Tai, J. C. & Carrasco, M. Covert attention affects the psychometric function of contrast sensitivity. *Vision Research* **42**, 949–967 (2002).

[23] Gorea, A., Caetta, F. & Sagi, D. Criteria interactions across visual attributes. *Vision Research* **45**, 2523–2532 (2005).

[24] Zak, I., Katkov, M., Gorea, A. & Sagi, D. Decision criteria in dual discrimination tasks estimated using external-noise methods. *Atten Percept Psychophys* **74**, 1042–1055 (2012).

[25] 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).

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

[27] Gorea, A. & Sagi, D. Natural extinction: A criterion shift phenomenon. *Visual Cognition* **9**, 913–936 (2002).

[28] 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`.

[29] Kontsevich, L. L., Chen, C.-C., Verghese, P. & Tyler, C. W. The unique criterion constraint: a false alarm? *Nature Neuroscience* **5**, 707–author reply 707–8 (2002).

[30] Rausch, M. & Zehetleitner, M. Visibility is not equivalent to confidence in a low contrast orientation discrimination task. *Front Psychol* **7**, 591 (2016).

[31] Zizlsperger, L., Sauvigny, T. & Haarmeier, T. Selective attention increases choice certainty in human decision making. *PLoS ONE* **7**, e41136 (2012).

[32] Zizlsperger, L., Sauvigny, T., Händel, B. & Haarmeier, T. Cortical representations of confidence in a visual perceptual decision. *Nat Commun* **5**, 3940 (2014).

[33] Wilimzig, C., Tsuchiya, N., Fahle, M., Einhäuser, W. & Koch, C. Spatial attention increases performance but not subjective confidence in a discrimination task. *JOV* **8**, 7–7 (2008).

[34] Ling, S. & Carrasco, M. Sustained and transient covert attention enhance the signal via different contrast response functions. *Vision Research* **46**, 1210–1220 (2006).

[35] Carrasco, M., Ling, S. & Read, S. Attention alters appearance. *Nature Neuroscience* **7**, 308–313 (2004).

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

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

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

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

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

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

[42] Acerbi, L., Wolpert, D. M. & Vijayakumar, S. Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Comput Biol* **8**, e1002771 (2012).

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

[44] Neal, R. M. Slice sampling. *Annals of statistics* **31**, 705–741 (2003).

[45] Vehtari, A., Gelman, A. & Gabry, J. Pareto Smoothed Importance Sampling 1507.02646v4 (2015). 1507.02646.

[46] 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).

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

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