Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis

Hsin-Hung Li[1] & Wei Ji Ma[1,2]

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

Correspondence: hsin.hung.li@nyu.edu (H-H. L.)

Abstract

Decision confidence reflects our ability to evaluate the quality of decisions and guides subsequent behaviors. Experiments on confidence reports have almost exclusively focused on two-alternative decision-making. In this realm, the leading theory is that confidence reflects the probability that a decision is correct (the posterior probability of the chosen option). There is, however, another possibility, namely that people are less confident if the *best two* options are closer to each other in posterior probability, regardless of how probable they are in *absolute* terms. This possibility has not previously been considered because in two-alternative decisions, it reduces to the leading theory. Here, we test this alternative theory in a three-alternative visual categorization task. We found that confidence reports are best explained by the difference between the posterior probabilities of the best and the next-best options, rather than by the posterior probability of the chosen (best) option alone, or by the overall uncertainty (entropy) of the posterior distribution. Our results upend the leading notion of decision confidence and instead suggest that confidence reflects the observer's subjective probability that they made the best possible decision.

Introduction

Confidence refers to the "sense of knowing" that comes with a decision. Confidence affects the planning of subsequent actions after a decision[1, 2], learning[3], and cooperation in group decision making[4]. Failures in utilizing confidence information have been linked to psychiatric disorders[5].

While human observers can report their self-assessment of the quality of their decisions[6, 7, 8, 9, 10, 11, 12], the computations underlying confidence reports are still insufficiently understood. The leading theory of confidence suggested that confidence reflects the probability that a decision is correct[7, 8, 13, 14, 15, 16, 17]. We refer to this idea as the "Bayesian confidence hypothesis" meaning that the decision-maker uses the posterior probability of the chosen category (i.e. the probability that decision is correct) for their confidence reports. In neurophysiological studies, a brain region or a neural process is considered to represent confidence if its responses correlate with the probability that a decision is correct[18, 19, 20]. Behavioral studies testing whether human confidence reports follow Bayesian confidence hypothesis have shown mixed results: While some studies found resemblances between Bayesian confidence and empirical data e.g. [18, 19, 21, 22], others have suggested that confidence reports deviate from the Bayesian confidence hypothesis e.g. [23, 24, 25].

Even though the Bayesian confidence hypothesis is the leading theory of confidence, there is currently no evidence to rule out the possibility that confidence is affected by unchosen options. Specifically, people could be less confident if the next-best option is very close to the best option. In other words, confidence could depend on the *difference* between the posterior probabilities of the best and the next-best options, rather than on the absolute value of the posterior of the best option. This idea has not been tested because previous studies of decision confidence have predominantly used two-alternative decision tasks; in such tasks, the alternative hypothesis is equivalent to the Bayesian confidence hypothesis, because the difference between the two posterior probabilities in a two-alternative task is a monotonic function of the highest posterior probability. Thus, to dissociate these two models of confidence, we need more than two alternatives. Therefore, we use a three-alternative decision task. To preview our main result, we find that the difference-based model accounts well for the data, whereas the model corresponding to the Bayesian confidence hypothesis and a third, entropy-based model do not.

Results

To investigate the computations underlying confidence reports in the presence of multiple alternatives, we designed a three-alternative categorization task. On each trial, participants viewed a large number of exemplar dots from each of the three categories (color-coded), along with one target dot in a different color (**Figure 1A**). Each category corresponded to an uncorrelated, circularly symmetric Gaussian distribution in the plane. We asked participants to regard the stimulus as a bird's eye view of three groups of people. People within a group wear shirts of the same color, and the target dot represents a person from one of the three groups. Participants made two responses: the category of the target, and their confidence in their decision on a four-point Likert scale.

To manipulate participants' beliefs (posterior probability distribution), we used different configurations of the category distributions and varied the position of the target dot within each configuration (**Figure 1B and 1C**). This design allowed us to test quantitative models of how the posterior distribution gives rise to confidence reports (see an illustration of this idea in **Supplementary Figure 1**).

Model

Generative model. Each category is equally probable. We assume that the observer makes a noisy measurement $\mathbf{x}$ of the position $\mathbf{s}$ of the target dot. We model the noise as obeying a circularly symmetric Gaussian distribution centered at the target dot.

Decision model. We now consider a Bayesian observer. We assume that the observer knows that each category is equally probable, and knows the distribution associated with each category (group) based on the exemplar dots. Given a measurement $\mathbf{x}$, the posterior probability of category C is then

$$p\left(C\middle|\mathbf{x}\right)=\frac{p\left(\mathbf{x}\middle|\mathbf{s}\right)p\left(\mathbf{s}\middle|C\right)}{\sum_{C=1}^{3}p\left(\mathbf{x}\middle|\mathbf{s}\right)p\left(\mathbf{s}\middle|C\right)}. \tag{1}$$

We further assume that due to decision noise or inference noise, the observer might not maintain the exact posterior distribution, $p(C|\mathbf{x})$, but instead a noisy version of it. This type of decision noise is consistent with the notion that a portion of variability in behavior is due to "late noise" at the level of decision variable[26, 27, 28]. We modeled decision noise by drawing a noisy posterior distribution from a Dirichlet distribution around the true posterior (**Figure 2A-B**; See details in **Methods**). In our case, the true posterior, which we denote by $\mathbf{p}$, consists of the three posterior probabilities from Eq.(1): $\mathbf{p}=(p(C=1|\mathbf{x}), p(C=2|\mathbf{x}), p(C=3|\mathbf{x}))$. The magnitude of the decision noise, the amount of variation around $\mathbf{p}$, is (inversely) controlled by a concentration parameter $\alpha>0$. When $\alpha\to\infty$, the variation vanishes and the posterior is noiseless. In general, the "noisy posterior", which we denote as a vector $\mathbf{p}_{\text{noisy}}$, satisfies

$$\mathbf{p}_{\text{noisy}} \sim \text{Dirichlet}\left(\alpha\mathbf{p}\right)$$

We assume that when reporting the category of the target, the observer chooses the category C with the highest $p_{\text{noisy}}(C|\mathbf{x})$. Unless otherwise specified, from now on we will refer to the noisy posterior distribution as simply the posterior distribution.

We introduce three models of confidence reports: the *Max* model, the *Entropy* model and the *Difference* model. Each of these models contains two steps: a) mapping the posterior distribution ($\mathbf{p}_{\text{noisy}}$) to a real-valued confidence variable; b) applying three criteria to this confidence variable to divide its space into four regions, which then map in increasing order to the four confidence ratings. The second step accounts for every possible monotonic mapping from the confidence variable to the four-point confidence rating. The three models differ in the first step.

The *Max* model corresponds to the Bayesian confidence hypothesis. In this model, the confidence variable is the probability that the chosen category is correct, or in other words, it is the highest of the three posterior probabilities (**Figure 2C**). In this model, the observer is least confident when the posterior distribution is uniform. Importantly, confidence is never influenced by the posterior probabilities of the categories that were not chosen.

In the *Difference* model, the confidence variable is the difference between the highest and second-highest posterior probabilities. In this model, confidence is low if the evidence for the next-best option is strong, and the observer is least confident whenever the two most probable categories are equally probable. One interpretation of this model is that confidence reflects the observer's subjective probability that they made the *best possible* choice, regardless of the actual posterior probability of that choice. An alternative interpretation is that decision-making consists of an iterative process in which the observer reduces a multiple-choice task to simpler (binary) choices (see Discussion).

In the *Entropy* model, the confidence variable is the negative of the uncertainty conveyed by the entire posterior distribution, quantified by its negative entropy. High confidence is associated with low entropy, and vice versa. Like in the Max model, the observer is least confident when the posterior distribution is uniform. Unlike in the Max model, however, the posterior probabilities of the non-chosen categories affect confidence. See the details of the models in **Methods**.

Note that all three models are Bayesian in a way that they compute the posterior probability distribution, and categorize the target dot by choosing the category with the highest posterior. The three models differ in how the confidence variable is read out from the posterior distribution. Only the Max model corresponds to the Bayesian confidence hypothesis. Only the Max model assumes that the posterior of the unchosen categories does not affect confidence. Importantly, in our three-alternative task, these models generate qualitatively different mappings from the posterior distribution to the confidence variable (**Figure 2C**). In a standard two-alternative task, however, the models would have been indistinguishable, because the probability of the non-chosen category would be determined by the probability of the chosen category.

We fitted the free parameters to the data of each individual subject using maximum-likelihood estimation, where the data on a given trial consist of a decision-confidence pair. Thus, we accounted for the joint distribution of decisions and confidence ratings[24, 25, 29] (see **Methods**). We compared models using the Akaike Information Criterion (AIC; Akaike, 1998). A model recovery analysis suggests that if the true model is among our tested models, our model comparison procedure is able to identify the correct model (see **Methods** and **Supplementary Figure 3**).

Experiment 1

In Experiment 1, the centers of the three category distributions were aligned vertically (**Figure 1B**). There were four conditions: In the first two conditions, the centers were evenly spaced horizontally. In the last two conditions, the center of the central distribution was closer to the center of either the left or the right distribution. The vertical position of the target dot was sampled from a normal distribution, and the horizontal position of the target dot was sampled uniformly between the center of the leftmost and right-most classes plus an extension to the left and the right (see **Methods**).

We plotted the psychometric curves (mean confidence rating as a function of the horizontal position of the target dot) by averaging confidence reports across trials using a sliding window (**Figure 3**). Mean confidence rating varied as a function of the horizontal position of the target. In the first two conditions (**Figure 3**), where the three distributions were evenly spaced, the psychometric curves showed two dips, with the lowest confidence attained at two positions symmetric around 0°.

We simulated the predicted psychometric curves using the best-fitting parameters of each model (**Figure 3B**). The fits of the Max and the Difference models resembled the data, but the best fit of the Entropy model showed a dip at the center in the first condition.

In the third and fourth conditions, in which the three distributions were unevenly spaced, mean confidence was lowest around the centers of the two distributions that were closest to each other. Only the Difference model exhibited this pattern, while the Max and the Entropy models deviated more clearly from the data.

The models not only make predictions for confidence ratings, but also for the category decisions (**Supplementary Figure 2**). Participants categorized the target dot based on its location, and when the target dot was close to the boundary between two categories (the location where two categories have equal likelihood), they assigned the target to those two categories with nearly equal probabilities. In general, this pattern is consistent with an observer who chooses the category associated with the highest posterior probability. The

Entropy model fits worst, even though all three models used the same rule for the category decision; this is because the confidence data also need to be accounted for.

Using the Akaike Information Criterion for model comparison (**Figure 4A and Supplementary Table 1**), we found that the Difference model outperformed the Max model by a group-averaged AIC score of 27.3 ± 7.0 (mean ± s.e.m.) and the Entropy model by 149 ± 25 (mean ± s.e.m.).

We further tested reduced versions of each of the three confidence models by removing either the sensory noise or the decision noise from the model. The Difference model outperformed the Max model and the Entropy model regardless of these manipulations (**Supplementary Figure 4 and Supplementary Table 1**). The sensory noise played a minor role in this task compared to the decision noise. For example, removing the sensory noise from the Difference model increased the AIC by 9.9 ± 3.2, while removing the inference noise increased the AIC by 57.3 ± 6.5. Using the Bayesian information criterion[30] for model comparison led to the same conclusions (**Supplementary Figure 5**).

Experiment 2

In Experiment 2, we aimed to test whether the findings in Experiment 1 could be generalized to other stimulus configurations, where the centers of the categories varied in a two-dimensional space. We tested four conditions in which the centers of the three groups varied along both horizontal and vertical axis (**Figure 1C**). We sampled the target dot positions uniformly within a circular area centered on the screen. In addition, the distribution of the categories used in Experiment 2 allowed us to probe confidence reports in a wider range of posterior distributions (**Supplementary Figure 1B**). For example, we can probe the confidence report when the target dot had the same distance to all three categories in Experiment 2, but not in Experiment 1.

The "psychometric curve" now is a heat map in two dimensions (**Figure 5**). The fits to these psychometric curves showed different patterns among the three models: When the three groups formed an equilateral triangle (**Figure 5,** the first and second columns), the confidence (as a function of target location) estimated by the Entropy model exhibited contours that were more convex than that in the data. In the last two conditions (**Figure 5,** the third and fourth columns), compared to the other two models, the Difference model showed stronger resemblance to the data, as the model exhibited an extended low confidence region at the side where two categories were positioned closely. The results of model comparisons were consistent with Experiment 1. The Difference model outperformed the Max model by a group-averaged AIC score of 45.9 ± 8.5 (mean ± s.e.m.) and the Entropy model by 152 ± 25 (mean ± s.e.m.) (**Figure 4B and Supplementary Table 1**). The model with both sensory and inference noise explained the data the best, and the inference noise had a stronger influence on the model fit than the sensory noise (**Supplementary Figure 4B, Supplementary Figure 5B and Supplementary Table 1**).

Experiment 3

So far, we found that the Difference model fits the data better than the Max and the Entropy. However, whether participants report the probability that a decision is correct (the Max model) might depend on the experimental design. In Experiment 1 and 2, participants received no feedback on their category decision. Thus, the probability of being correct in the task could be difficult to learn. To investigate this issue, in Experiment 3, using the same four stimulus configurations as those in Experiment 1 (**Figure 1B**), we randomly chose one of the three groups as the true target category in each trial, and sampled the target position from the distribution of the true category. Feedback was presented at the end of each trial, informing participants of the true category.

The results of model comparison were consistent with Experiment 1. The Difference model outperformed the Max model by a group-averaged AIC score of 10.3 ± 2.9 (mean ± s.e.m.) and the Entropy model by 93 ± 18 (mean ± s.e.m.) (**Supplementary Figure 6 and Supplementary Table 1**). The model with both sensory and inference noise explained the data the best, and the inference noise had a stronger influence on the model fit than the sensory noise (**Supplementary Figure 4C and 5C**).

Discussion

To distinguish the leading model of perceptual confidence (the Bayesian confidence hypothesis) from a new alternative model in which confidence is affected by the posterior probabilities of unchosen options, we studied human confidence reports in a three-alternative perceptual decision task. We found that confidence is best described by the Difference model, in which confidence reflects the difference between the strength of observers' belief (posterior probability) of the top two options in a decision. The Max model (which corresponds to the Bayesian confidence hypothesis) and the Entropy model (in which confidence is derived from the entropy of the posterior distribution) fell short in accounting for the data. Our results were robust under changes of stimulus configurations (Experiment 1 and 2), and when trial-by-trial feedback was provided (Experiment 3). Our results demonstrate that the posterior probabilities of the unchosen categories impact confidence in decision-making.

Decision tasks with multiple alternatives not only allow us to dissociate different computational models of confidence, they are also ecologically important. In the real world, human and other animals often face decisions with multiple alternatives, such as identifying the color of a traffic light, recognizing a person, categorizing a species of an animal, online shopping, or making a medical diagnosis.

Our models can be generalized to categorical choice with more than three alternatives. Specifically, the Difference model predicts that besides the posterior probabilities of the top two options, the posterior of the other options does not matter as long as they add up to the same total. A special type of categorical choice is when the world state variable is continuous (e.g. in an orientation estimation task) but gets discretized for the purpose of the experiment. Consider the specific case that the posterior distribution is Gaussian. An observer following the Difference model would compute the difference between the posteriors of the two discrete options closest to the peak. This serves as a very coarse approximation to the curvature of the posterior distribution at its peak, which, for Gaussians, is monotonically related to its inverse variance, consistent with an earlier model in which confidence is based on the precision parameter of the posterior[29]. Outside the realm of Gaussian and similar distributions, the Difference model and van den Berg et al.'s model (2017) might be distinguishable. For example, when the posterior distribution is bimodal, with the modes slightly different in height, the variance of the posterior is dominated by the separation between the modes, whereas the Difference model will use the difference in height for confidence reports.

Although many behavioral studies have emphasized similarities between human confidence reports and predictions of Bayesian models e.g. [18, 19, 21, 22], the Bayesian confidence hypothesis has been questioned before[8, 13, 14, 15, 16]. In addition to the probability of being correct, confidence is influenced by various factors such as reaction time[31], post-decision processing[32, 33, 34, 35], and the magnitude of positive evidence[36, 37, 38, 39]. Two model comparison studies have shown deviations from Bayesian confidence hypothesis in two-alternative decision tasks[24, 25]. However, in one study[24], the experimental design did not allow the authors to strongly distinguish the model that was based on Bayesian confidence hypothesis from those that were not. Moreover, in both studies[24, 25], the alternative models were based on heuristic decision rules without a broader theoretical interpretation. Here, we have identified a type of deviation from the Bayesian predictions that is not only of a qualitatively different nature, but that also raises new theoretical questions.

Specifically, the Difference model is currently a descriptive model. We have two suggestions to interpret it as an outcome of approximate inference. First, the Difference model might be an approximation to a model in which confidence depends on the probability that an observer made the *best possible* decision. Specifically, the observer is "aware" that their decision is based on the noisy posterior $\mathbf{p}_{\text{noisy}}$ rather than the true posterior $\mathbf{p}$. Thus, it is possible that the chosen category is not the category with the highest probability in the true posterior. Confidence would be derived from the probability that the chosen category has the highest probability in the true posterior distribution. The observer achieves this computation using the evidence for the next-best option: The stronger the evidence for the next-best option, the more likely that the chosen category is not the top choice in the true posterior, thus leading to lower confidence. Recent work has shown that subjective confidence guides information seeking during decision-making[40]. Under the Difference model, during information seeking, the observer's goal is to make sure that the best option is better than the alternative options. Low confidence would encourage the observer to collect more information in order to strengthen the belief that the best option is better than the next-best option.

Second, the finding that confidence is best described by the relative strength of the evidence of the top two options might be related to other findings in multiple-alternative decision-making. For example, in one experiment, observers watched columns of bricks build up on the screen, and reported which column had the highest accumulation rate[41]. A heuristic model in which the observer makes a decision when the height of the tallest column exceeds the height of the next-tallest column by a fixed threshold captured the overall pattern of people's behavior. In a study on self-directed learning in a three-alternative categorization task, observers had to learn the category distributions by sampling from the feature space and receiving feedback. Instead of choosing the most informative samples, human observers chose ones for which the likelihood of two categories were similar, namely those located at boundaries between pairs of two categories[42]. This literature allows us to speculate that observers might decompose a multiple-alternative decision into several simpler (perhaps binary) choices. This notion is reminiscent of the concept in prospect theory that before a phase of evaluation, extremely unlikely outcomes might be first discarded in an "editing" phase[43]. Hence, an alternative interpretation of our results is that confidence reports deviate from the Bayesian confidence hypothesis (the Max model) because the observer estimates the probability of correct in a way that ignores the options that are discarded before final evaluation. In the Difference model, the least favorite option is not completely discarded because it decreases the posterior probabilities of the other two options (and thus their difference) by contributing to the normalization pool[44, 45]. Therefore, we consider an extreme version of editing, the Ratio model, in which the least-favorite option does not even participate in normalization, and thus confidence solely depends on the likelihood ratio between the top two options. The Difference model and the Ratio model are not distinguishable in Experiment 1 and 2 (**Supplementary Figure 7**). In Experiment 3, the Difference model was very similar to the Ratio model in group-averaged AIC (3.8 ± 1.4 in favor of the Difference model). Testing variable numbers of categories within an experiment might help to differentiate between these two models.

We found that compared to the sensory noise, the noise associated with the computation of posterior probability plays a more important role in our task. This is consistent with the findings of a recent study[26]. The relative unimportance of sensory noise could be partly due to our experimental designs, which used stimuli with strong signal strength (saturated color and unlimited duration). Different from our study, Drugowitsch et al. (2016) devised an evidence accumulation task and further distinguished two types of decision noises: First, the inference noise that was added (and thus increased) with each new stimulus sample. Second, the selection noise that was injected only once at the final response. Because our experiment only had one stimulus in each trial, these two sources of variability were indistinguishable.

Do our results generalize beyond perceptual decision-making? In a two-alternative value-based decision task, observers reported confidence in a way that was similar to that in perceptual decision tasks[10]: When observers were asked to choose the good with the higher value, confidence increased with the posterior probability that a decision is correct, which in turn increased with the difference in value between the two goods. In addition, choice accuracy was higher in high-confidence trials then in low-confidence trials, reflecting observers' ability to evaluate their own performance. It is unknown how observers compute confidence when there are more than two goods. In three-alternative value-based tasks, the Difference model would predict that, confidence is determined by the difference between the probability that the chosen item is the most valuable and the probability that the next-best item is the most valuable.

How does the present study advance our understanding of the neural basis of confidence? Most neurophysiological studies of confidence have considered the neural activity that correlates with the probability of being correct as the neural representation of confidence (but see [48]). Neural responses in parietal cortex[19], orbitofrontal cortex[18] and pulvinar[20] have been associated with that representation of confidence.. These studies all used two-alternative decision tasks. Multiple-alternative decision tasks have been used in neurophysiological studies on non-human primates but not with the objective of studying confidence[45, 49, 50, 51]. By utilizing multiple-alternative tasks, neural studies could dissociate the neural correlates of probability correct from that of the "difference" confidence variable in the Difference model, which according to our results might be the basis of human subjective confidence. A potentially important difference between human and non-human animal studies is that in the latter, confidence is not explicitly reported but operationalized through some aspect of behavior, such as the probability of choosing a "safe" (opt-out) option[19, 20, 46, 47, 48], or the time spent on waiting for reward[18]. Thus, one should be careful when directly comparing these implicit reports with explicit confidence reports in human studies.

Methods

Setup

Participants sat in a dimly lit room with the chin rest positioned 45 cm from the monitor. The stimuli and the experiment were controlled by customized programs written in Javascript.

The monitor had a resolution of 3840 by 2160 pixels and a refresh rate of 30 Hz. The spectrum and the luminance of the monitor were measured with a spectroradiometer.

Participants

Thirteen participants took part in Experiment 1. Eleven participants took part in Experiment 2. Eleven participants took part in Experiment 3. All participants had normal or corrected-to-normal vision. The experiments were conducted with the written consent of each participant. The University Committee on Activities involving Human Subjects at New York University approved the experimental protocols.

Stimulus

On each trial, three categories of exemplar dots (375 dots per category) were presented along with one target dot, a black dot (**Figure 1A**). The dots within a category were distributed as an uncorrelated, circularly symmetric Gaussian distribution with a standard deviation of 2° (degree visual angle) along both horizontal and vertical directions. Exemplar dots from the different categories were coded with different colors. The three colors were randomly chosen on each trial, and were equally spaced in Commission Internationale de l'Eclairage (CIE) L*a*b* color space. The three colors were at a fixed lightness of L*=70 and were equidistant from the gray point (a*=0, and b*=0).

In Experiment 1 and 3, the centers of the three categories were aligned vertically to the center of the screen, and were located at different horizontal positions (**Figure 1B**). In four configurations, the horizontal positions of the centers of the three categories were (-3°, 0°, 3°), (-4°, 0°, 4°), (-3°, -2°, 3°), and (-3°, 2°, 3°), from the center of the screen respectively. In Experiment 2, the centers of the three categories varied on a 2-dimensional space (**Figure 1C**). In four configurations, the horizontal positions of the centers of the three categories were (-2°, 0°, 2°), (-1.59°, 0°, 1.59°), (-2°, -2°, 2°), and (-2°, 2°, 2°), from the center of the screen, respectively. The vertical positions of the centers were (1.16°, -2.31°, 1.16°), (0.94°, -1.84°, 0.94°), (1.16°, 0°, 1.16°), (1.16°, 0°, 1.16°) from the center of the screen respectively.

Procedures

We told participants that the three groups of exemplar dots represented a bird's eye view of three groups of people. The three groups contained equal numbers of people. The black dot (the target) is a person from one of the three groups, but we do not know the color of her/his T-shirt. We asked participants to categorize the target to one of the three groups based on the (position) information conveyed by the dots, and report their confidence on a four-point Likert scale.

Each trial started with the onset of the stimulus and three rectangular buttons positioned at the bottom of the screen (**Figure 1A**). On each trial, participants first categorized the target to one of the three groups (based on the position information conveyed by the dots) by using the mouse to click on one of the three buttons. After participants reported their decision, the three buttons were replaced by four buttons (labeled as "very unconfident", "somewhat unconfident", "somewhat confident", and "very confident") for participants to report their confidence on the decision they made. The stimuli were presented throughout each trial. Reaction time (for both decision and confidence reports) was unlimited. After participants reported their confidence, all the exemplar dots and the rectangular buttons disappeared from the screen, and the next trial started after a 600 ms inter-trial-interval.

In Experiment 1, the vertical position of the target dot was sampled from a normal distribution (2° std), and the horizontal position of the target dot was sampled uniformly between the center of the leftmost and rightmost categories plus a 0.2° extension to the left and the right. In Experiment 2, the target dot was uniformly sampled from a circular area (2.6° radius) positioned at the center of the screen. No feedback was provided in Experiment 1 and Experiment 2.

In Experiment 3, in each trial, we randomly chose one of the three categories with equal probability as the true category. We then positioned the target dot by sampling from the distribution of the true category. A feedback regarding the true category was provided at the end of each trial: After participants reported their confidence, all exemplar dots disappeared except that the exemplar dots from the true category remained on the screen for an extra 500 ms. In each experiment, participants completed one 1-hr session (84 trials per configuration in Experiment 1 and 120 trials per configuration in Experiment 2 and 3). All the trials in one session were separated into 8 blocks with equal number of trials. Different configurations were randomized and interleaved within each block.

Models

Generative model. The target belongs to category $C \in \{1, 2, 3\}$. The two-dimensional position $\mathbf{s}$ of a target in category C is drawn from a two-dimensional Gaussian $p(\mathbf{s}|C) = N(\mathbf{s}; \mathbf{m}_C, \sigma_s^2\mathbf{I})$, where $\mathbf{m}_C$ is the center of category C, σ_s^2 is the variance of the stimulus distribution, and $\mathbf{I}$ is the 2-dimensional identity matrix. We assume that the observer make a noisy sensory measurement $\mathbf{x}$ of the target position. We model the sensory noisy using a Gaussian distribution centered at $\mathbf{s}$ with covariance matrix $\sigma^2\mathbf{I}$. Thus, the distribution of $\mathbf{x}$ given category C is $p(\mathbf{x}|C) = N(\mathbf{x}; \mathbf{m}_C, (\sigma_s^2+\sigma^2)\mathbf{I})$.

Inference on a given trial. We assume that the observer knows the mean and standard deviation of each category based on the exemplar dots, and that the observer assumes that the three categories have equal probabilities. The posterior probability of category C given the measurement $\mathbf{x}$ is then $p(C|\mathbf{x}) \propto p(\mathbf{x}|C) = N(\mathbf{x}; \mathbf{m}_C, (\sigma_s^2+\sigma^2)\mathbf{I})$. Instead of the true posterior $p(C|\mathbf{x})$, the observer makes the decisions based on $p_{\text{noisy}}(C|\mathbf{x})$, a noisy version of the posterior probability. We obtain a noisy posterior $p_{\text{noisy}}(C|\mathbf{x})$ by drawing from a Dirichlet distribution. The Dirichlet distribution is a generalization of the beta distribution. Just like the beta distribution is a continuous distribution over the probability parameter of a Bernoulli random variable, the Dirichlet distribution is a distribution over a vector that represents the probabilities of any number of categories. The Dirichlet distribution is parameterized as

$$p(\mathbf{p}_{\text{noisy}} \mid \mathbf{p};\alpha) = \frac{1}{B(\alpha\mathbf{p})}\prod_{i=1}^{3} p_{ni}^{\alpha p_i - 1}$$

$$B(\alpha\mathbf{p}) = \frac{\prod_{i=1}^{3}\Gamma(\alpha p_i)}{\Gamma(\sum_{i=1}^{3}(\alpha p_i))}$$

$\mathbf{p}$ is a vector consists of the three posterior probabilities, $\mathbf{p}=(p(C=1|\mathbf{x}), p(C=2|\mathbf{x}), p(C=3|\mathbf{x}))$. $\mathbf{p}_{\text{noisy}}$ is a vector consists of the three posterior probabilities perturbed by the decision noise, $\mathbf{p}_{\text{noisy}} =(p_{\text{noisy}}(C=1|\mathbf{x}), p_{\text{noisy}}(C=2|\mathbf{x}), p_{\text{noisy}}(C=3|\mathbf{x}))$. The mean of $p_{\text{noisy}}(C|\mathbf{x})$ is $p(C|\mathbf{x})$. The concentration parameter α inversely determines the magnitude of the decision noise. To make a category decision, the observer chooses the category that maximizes the posterior probability: $\hat{C} = \underset{C}{\operatorname{argmax}}\, p_{\text{noisy}}\left(C \mid \mathbf{x}\right)$.

We considered three models of confidence reports. We first specify in each model an internal continuous confidence variable c^*. In the *Max* (maximum a posteriori) model, c^* is the posterior probability of the chosen category: $c^* = p_{\text{noisy}}\left(C = \hat{C} \mid \mathbf{x}\right)$. In the Difference model, c^* is a difference: $c^* = p_{\text{noisy}}\left(C = \hat{C} \mid \mathbf{x}\right) - p_{\text{noisy}}\left(C = \hat{C}_2 \mid \mathbf{x}\right)$, where $\hat{C}_2$ is the category with the second-highest posterior probability. In the *Entropy* model, c^* is the negative entropy of the posterior distribution: $c^* = \sum_{C=1}^{3} p_{\text{noisy}}\left(C \middle| \mathbf{x}\right) \log p_{\text{noisy}}\left(C \middle| \mathbf{x}\right)$.

In each model, the continuous confidence variable c^* is converted to a four-point confidence report c by imposing three confidence criteria b_1, b_2 and b_3. For example, c=3 when $b_2 < c^* < b_3$. We also included a lapse rate λ in each model; on a lapse trial, the observer presses a random button for both the decision and the confidence report. In addition to the models that included both sensory and decision noise, we took a factorial approach and tested various combinations of confidence model and sources of variability [52, 53, 54]. For each confidence model, we tested two reduced models by removing either the sensory noise (by setting σ=0) or the decision noise (by setting $p_{\text{noisy}}(C|\mathbf{x}) = p(C|\mathbf{x})$) from the model.

Response probabilities. So far, we have described the mapping from a measurement $\mathbf{x}$ to a decision $\hat{C}$ and a confidence report c. The measurement, however, is internal to the observer and unknown to the experimenter. Therefore, to obtain model predictions for a given parameter combination (σ, α, b_1, b_2, b_3, λ), we perform a Monte Carlo simulation. For every true target position $\mathbf{s}$ that occurs in the experiment**,** we simulated a large number (10,000) of measurements $\mathbf{x}$. For each of these measurements, we compute the posterior $p(C|\mathbf{x})$, add decision noise to obtain $p_{\text{noisy}}(C|\mathbf{x})$, and finally obtain a category decision $\hat{C}$ and a confidence report c. Across all simulated measurements, we obtain a joint distribution $p\left(\hat{C}, c \middle| \mathbf{s}; \sigma, \alpha, b_1, b_2, b_3, \lambda\right)$ that represents the response probabilities of the observer.

Model fitting and model comparison. We denote the parameters (σ, α, b_1, b_2, b_3, λ) collectively by θ. We fit each model to individual-subject data by maximizing the log likelihood of θ, log L(θ)=log p(data|θ). We assume that the trials are conditionally independent. We denote the target position, category response, and four-point confidence report on the ith trial by $\mathbf{s}_i$, $\hat{C}_i$, and c_i, respectively. Then, the log likelihood becomes

$$\log L(\theta) = \log \prod_i p\left(\hat{C}_i, c_i \middle| \mathbf{s}_i, \theta\right) = \sum_i \log p\left(\hat{C}_i, c_i \middle| \mathbf{s}_i, \theta\right),$$

where $p\left(\hat{C}_i, c_i \middle| \mathbf{s}_i, \theta\right)$ is obtained from the Monte Carlo simulation described above. We optimized the parameters using a new method called Bayesian Adaptive Direct Search [55]. We used AIC and BIC for model comparison. To report the AIC (or BIC) index, we computed the AIC (or BIC) for each individual and then averaged the AIC across participants.

Parameterization

The full version of the three confidence models (Max, Difference and Entropy models reported in **Figure 4**) have the same set of free parameters including the magnitude of sensory noise (σ), the magnitude (concentration parameter) of decision noise (α), three boundaries for converting continuous confidence variable to button press (b_1, b_2, b_3) and a lapse rate λ.

For each of the three confidence models, we tested two versions of the reduced models (reported in **Supplementary Figure 4** and **Supplementary Figure 5**). In one version, we kept the sensory noise (σ) in the model while removing the decision noise (α). In the other version we kept the decision noise (α) in the model while removing the sensory noise (σ).

Model Recovery

To evaluate our ability to distinguish the three models, we performed a model recovery analysis. Based on the design of Experiment 1, we synthesized 10 datasets for each of the confidence models. To ensure that the synthesized data resemble our experimental data, we synthesized the data using the group-averaged best-fitting parameter values obtained in Experiment 1. We then fit each of the 30 datasets (3 generating models with 10 datasets each) with the 3 models. Supplementary Figure 3 illustrates the results averaged over 10 datasets for each of the generating model.

Data visualization

For Experiment 1 and 3, we used a sliding window to visualize the psychometric curves, defined as the confidence ratings as a function of horizontal location of the target dot. The sliding window had a width of 0.6°. We moved the window horizontally (in a step of 0.1°) from the left to the right of the screen center. At each step, we computed mean confidence rating by averaging the confidence reports *c* of all the trials fell within the window (based on the horizontal target location of each trial). We first applied this procedure to individual data, and then averaged the individual psychometric curves across subjects (the black curves in **Figure 3B** and **Supplementary Figure 6B)**. For Experiment 1, we visualized the data ranging from -3.5° to +3.5° from the screen center. For Experiment 3, we visualized the data ranging from -5° to +5° from the center. These ranges were chosen so that each steps along the black curves in **Figure 3B** and **Supplementary Figure 6B** contained at least 5 trials per subject on average. To visualize the model fit, we sampled a series of target dot locations along the horizontal axis (in a step of 0.1°), and we used the best-fitting parameters to compute the confidence rating predicted by the models for each target location. We then used the same procedure (a sliding window) to compute the mean confidence rating predicted by the models (the blue curves in **Figure 3B** and **Supplementary Figure 6B**).

For Experiment 2, the "psychometric curve" became a heat map in a two-dimensional space (**Figure 5**). We tiled the two-dimensional space with non-overlapped hexagonal spatial windows (with a radius of 0.25°) positioned from -3° to +3° (**Figure 5A**) along both horizontal and vertical axis. To compute the mean confidence rating for each hexagonal window, we averaged the confidence ratings across all the trials fell within that window for each participant. If the number of trials was zero among all the participants for a window, that window was left as white in **Figure 5A**. To visualize the model fit, we used the best-fitting parameters and computed the confidence rating predicted by the models for an array of target locations (a grid tiling the two-dimensional space with a step of 0.1° along both horizontal and vertical axis). The predicted confidence rating was then averaged within each hexagonal window.

Acknowledgement

We thank members of the Ma Lab, Hui-Kuan Chung, Rachel Denison, and Michael Landy for helpful comments on the manuscript.

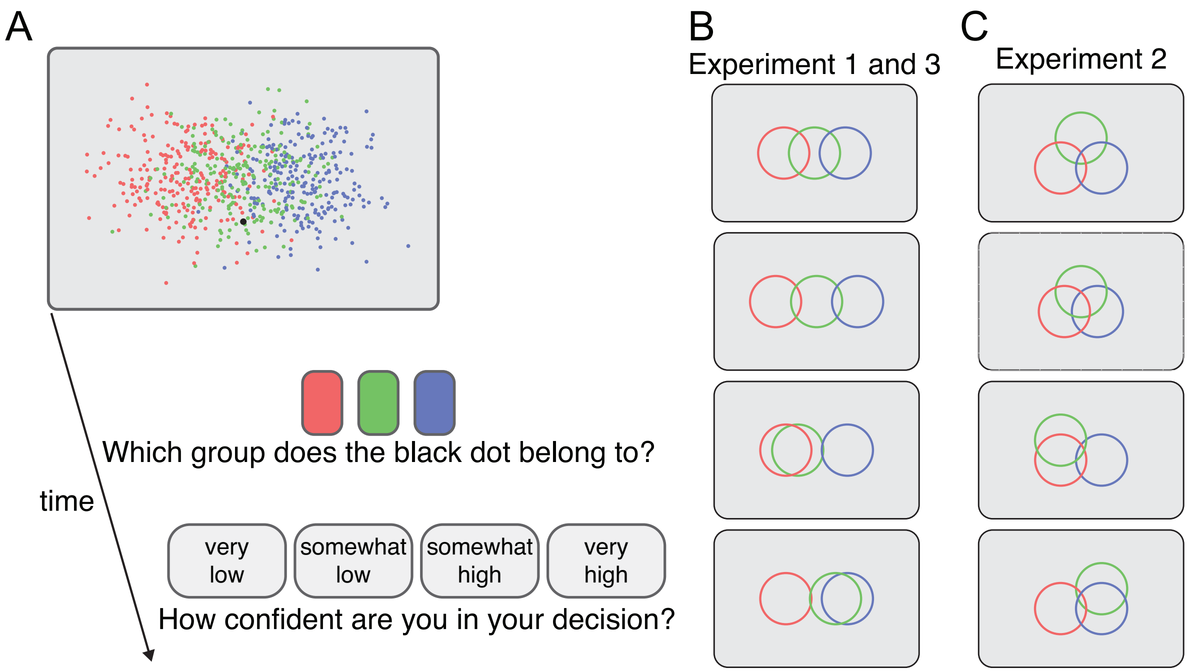


Figure 1. (A) Experimental procedure. Each trial started with the presentation of the stimulus including exemplar dots in three different colors representing the distribution of each of the three categories and one target dot, the black dot. Observers first reported their decisions in the categorization task and then reported their confidence by using the rectangular buttons presented at the bottom of the screen. (B) and (C) Schematic representation of the distribution of the categories. The circles are centered at the mean location of each category. The width of the circles corresponds to 2.5 times the standard deviation of the category distribution. (B) The four conditions tested in Experiment 1 and 3. (C) The four conditions tested in Experiment 2. The exemplar dots in (A) are based on the distribution depicted in the top panel in (B).

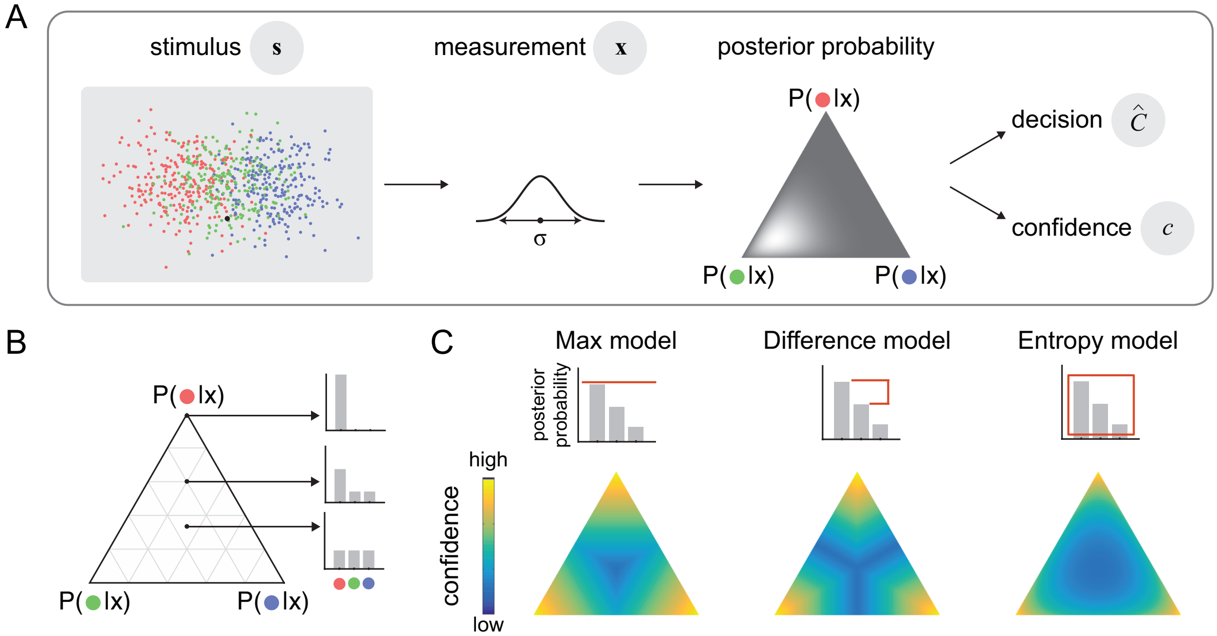


Figure 2. (A) Generative model. Target position is represented by s. Two sources of variability are considered in the model: First, observers have access to noisy measurement x, a Gaussian distribution centered at s with a standard deviation σ. Second, given the same measurement x, the posterior distribution varies across trials due to decision noise, modeled by Dirichlet distribution, of which spread (represented by the shade of the ternary plot) is controlled by a parameter α (see Methods). On each trial, a decision $\hat{C}$ and a confidence c are read out from the posterior distribution of that trial. (B) We use ternary plots to represent all possible posterior distributions. For example, a point at the center represents a uniform posterior distribution; at the corners of the ternary plot, the posterior probability of one category is one while the posterior for the other two categories are zeros. (C) The bar graphs illustrate how confidence is read out from posterior probabilities in each model. The color of each ternary plot represents the confidence as a function of posterior distribution for each model. The color is scaled for each ternary plot (independently) to take the whole range of the color bar.

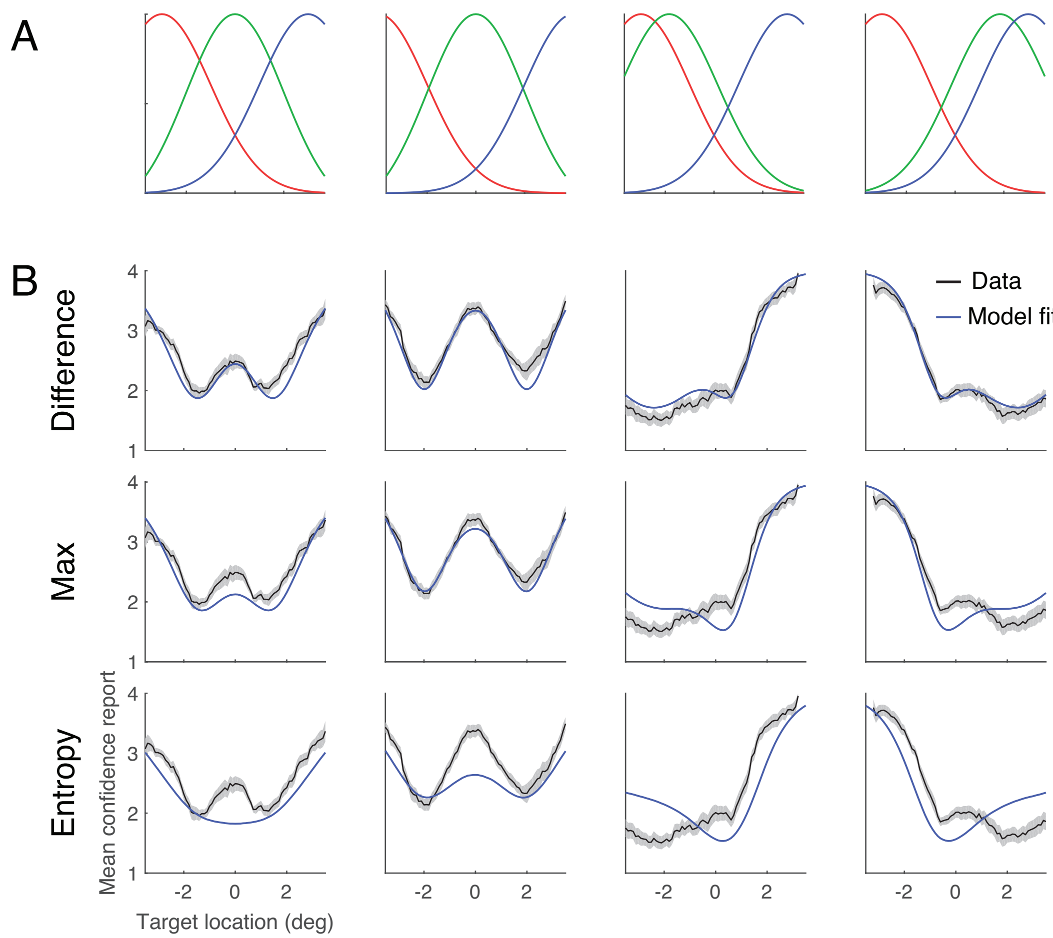


Figure 3. Experiment 1. (A) The distribution of the reference dots in each condition. (B) Mean confidence rating as a function of target position for each of the four conditions. The black curves represent group mean ± 1 s.e.m. Blue curves represent the model fit averaged across individuals.

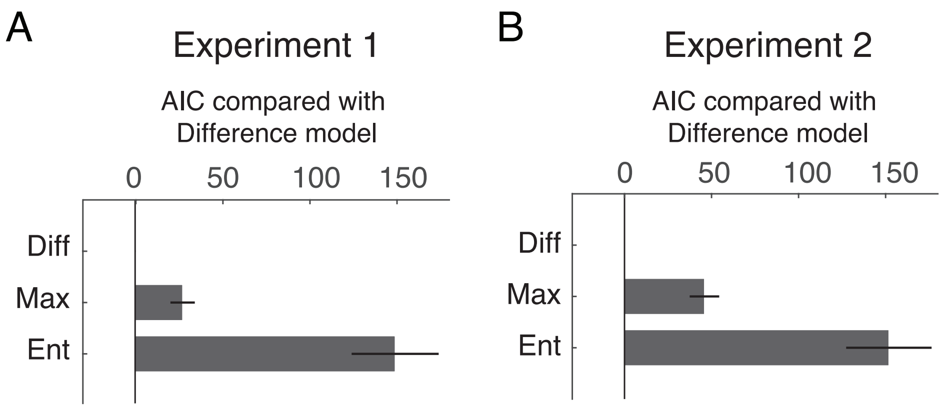


Figure 4. Model comparisons using ΔAIC: AIC of each model compared with the Difference model. The bars represent ΔAIC averaged across participants. The error bars represent ± 1 s.e.m across participants. (A) Experiment 1. (B) Experiment 2.

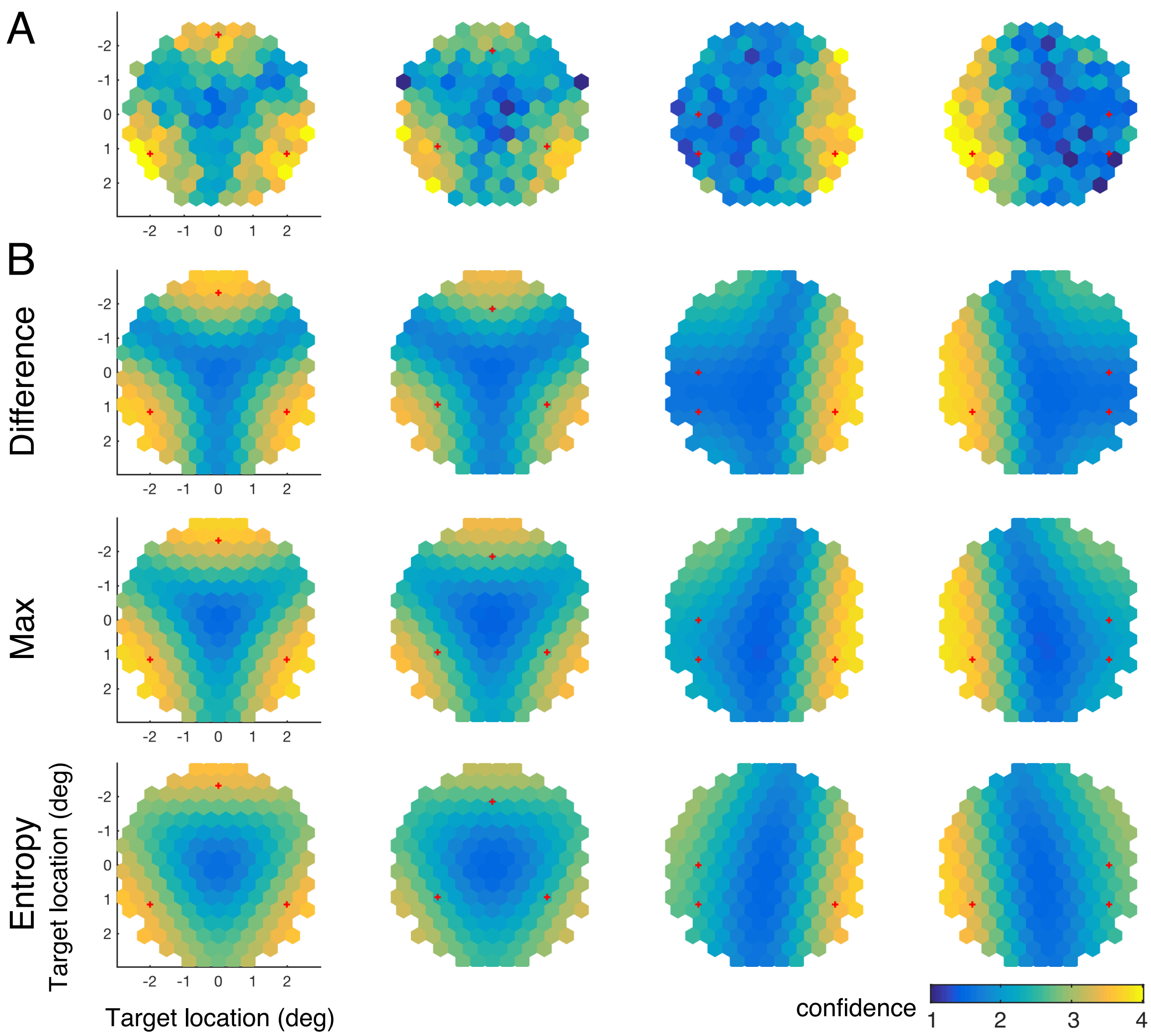


Figure 5. Experiment 2. (A) The mean confidence rating as a function of target positions. (B) Model fit averaged across individuals. The red crosses in each panel represent the center of each of the three categories.

References

1. Persaud N, McLeod P, Cowey A. Post-decision wagering objectively measures awareness. *Nature neuroscience* **10**, 257 (2007).
2. Van den Berg R, Zylberberg A, Kiani R, Shadlen MN, Wolpert DM. Confidence Is the Bridge between Multi-stage Decisions. *Current Biology* **26**, 3157-3168 (2016).
3. Meyniel F, Schlunegger D, Dehaene S. The sense of confidence during probabilistic learning: A normative account. *PLoS computational biology* **11**, e1004305 (2015).
4. Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD. Optimally interacting minds. *Science* **329**, 1081-1085 (2010).
5. Vaghi MM, Luyckx F, Sule A, Fineberg NA, Robbins TW, De Martino B. Compulsivity Reveals a Novel Dissociation between Action and Confidence. *Neuron* **96**, 348-354. e344 (2017).
6. Fleming SM, Lau HC. How to measure metacognition. *Frontiers in human neuroscience* **8**, 443 (2014).
7. Mamassian P. Visual Confidence. *Annual Review of Vision Science* **2**, 459-481 (2016).
8. Kepecs A, Mainen ZF. A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **367**, 1322-1337 (2012).
9. Yeung N, Summerfield C. Metacognition in human decision-making: confidence and error monitoring. *Phil Trans R Soc B* **367**, 1310-1321 (2012).
10. De Martino B, Fleming SM, Garrett N, Dolan RJ. Confidence in value-based choice. *Nature neuroscience* **16**, 105 (2013).
11. Lebreton M, Abitbol R, Daunizeau J, Pessiglione M. Automatic integration of confidence in the brain valuation signal. *Nature neuroscience* **18**, 1159 (2015).
12. Polania R, Woodford M, Ruff CC. Efficient coding of subjective value. *Nature neuroscience* **22**, 134 (2019).
13. Pouget A, Drugowitsch J, Kepecs A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience* **19**, 366 (2016).

14. Drugowitsch J, Moreno-Bote R, Pouget A. Relation between belief and performance in perceptual decision making. *PloS one* **9**, e96511 (2014).
15. Clarke FR, Birdsall TG, Tanner Jr WP. Two types of ROC curves and definitions of parameters. *The Journal of the Acoustical Society of America* **31**, 629-630 (1959).
16. Galvin SJ, Podd JV, Drga V, Whitmore J. Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review* **10**, 843-876 (2003).
17. Peirce CS, Jastrow J. On small differences in sensation. (1884).
18. Kepecs A, Uchida N, Zariwala HA, Mainen ZF. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227 (2008).
19. Kiani R, Shadlen MN. Representation of confidence associated with a decision by neurons in the parietal cortex. *science* **324**, 759-764 (2009).
20. Komura Y, Nikkuni A, Hirashima N, Uetake T, Miyamoto A. Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature neuroscience* **16**, 749 (2013).
21. Sanders JI, Hangya B, Kepecs A. Signatures of a statistical computation in the human sense of confidence. *Neuron* **90**, 499-506 (2016).
22. Barthelmé S, Mamassian P. Flexible mechanisms underlie the evaluation of visual confidence. *Proceedings of the National Academy of Sciences* **107**, 20834-20839 (2010).
23. Navajas J, Hindocha C, Foda H, Keramati M, Latham PE, Bahrami B. The idiosyncratic nature of confidence. *Nature human behaviour* **1**, 810 (2017).
24. Aitchison L, Bang D, Bahrami B, Latham PE. Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS computational biology* **11**, e1004519 (2015).
25. Adler WT, Ma WJ. Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLOS Computational Biology* **14**, e1006572 (2018).
26. Drugowitsch J, Wyart V, Devauchelle A-D, Koechlin E. Computational precision of mental inference as critical source of human choice suboptimality. *Neuron* **92**, 1398-1411 (2016).

27. Keshvari S, Van den Berg R, Ma WJ. Probabilistic computation in human perception under variability in encoding precision. *PLoS One* **7**, e40216 (2012).
28. Shen S, Ma WJ. Variable precision in visual perception. *Psychological Review* **126**, 89-132 (2019).
29. van den Berg R, Yoo AH, Ma WJ. Fechner's law in metacognition: A quantitative model of visual working memory confidence. *Psychological review* **124**, 197 (2017).
30. Schwarz G. Estimating the dimension of a model. *The annals of statistics* **6**, 461-464 (1978).
31. Kiani R, Corthell L, Shadlen MN. Choice certainty is informed by both evidence and decision time. *Neuron* **84**, 1329-1342 (2014).
32. Moran R, Teodorescu AR, Usher M. Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive psychology* **78**, 99-147 (2015).
33. Pleskac TJ, Busemeyer JR. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological review* **117**, 864 (2010).
34. Yu S, Pleskac TJ, Zeigenfuse MD. Dynamics of postdecisional processing of confidence. *Journal of Experimental Psychology: General* **144**, 489 (2015).
35. Navajas J, Bahrami B, Latham PE. Post-decisional accounts of biases in confidence. *Current Opinion in Behavioral Sciences* **11**, 55-60 (2016).
36. Koizumi A, Maniscalco B, Lau H. Does perceptual confidence facilitate cognitive control? *Attention, Perception, & Psychophysics* **77**, 1295-1306 (2015).
37. Zylberberg A, Barttfeld P, Sigman M. The construction of confidence in a perceptual decision. *Front Integr Neurosci* **6**, 2359-2374 (2012).
38. Peters MA*, et al.* Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature human behaviour* **1**, 0139 (2017).
39. Maniscalco B, Lau H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition* **21**, 422-430 (2012).
40. Desender K, Boldt A, Yeung N. Subjective confidence predicts information seeking in decision making. *Psychological science* **29**, 761-778 (2018).

41. Brown S, Steyvers M, Wagenmakers E-J. Observing evidence accumulation during multi-alternative decisions. *Journal of Mathematical Psychology* **53**, 453-462 (2009).

42. Markant DB, Settles B, Gureckis TM. Self‐directed learning favors local, rather than global, uncertainty. *Cognitive science* **40**, 100-120 (2016).

43. Kahneman D, Tversky A. Prospect theory: An analysis of decision under risk. In: *Handbook of the fundamentals of financial decision making: Part I* (ed^(eds). World Scientific (2013).

44. Carandini M, Heeger DJ. Normalization as a canonical neural computation. *Nature Reviews Neuroscience* **13**, 51-62 (2012).

45. Louie K, Grattan LE, Glimcher PW. Reward value-based gain control: divisive normalization in parietal cortex. *Journal of Neuroscience* **31**, 10627-10639 (2011).

46. Hampton RR. Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences* **98**, 5359-5362 (2001).

47. Foote AL, Crystal JD. Metacognition in the rat. *Current Biology* **17**, 551-555 (2007).

48. Odegaard B, Grimaldi P, Cho SH, Peters MA, Lau H, Basso MA. Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proceedings of the National Academy of Sciences*, 201711628 (2018).

49. Churchland AK, Kiani R, Shadlen MN. Decision-making with multiple alternatives. *Nature neuroscience* **11**, 693 (2008).

50. Churchland AK, Ditterich J. New advances in understanding decisions among multiple alternatives. *Current opinion in neurobiology* **22**, 920-926 (2012).

51. Ditterich J. A comparison between mechanisms of multi-alternative perceptual decision making: ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. *Frontiers in neuroscience* **4**, 184 (2010).

52. Acerbi L, Wolpert DM, Vijayakumar S. Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS computational biology* **8**, e1002771 (2012).

53. van den Berg R, Awh E, Ma WJ. Factorial comparison of working memory models. *Psychological review* **121**, 124 (2014).
54. Daunizeau J, Preuschoff K, Friston K, Stephan K. Optimizing experimental design for comparing models of brain function. *PLoS computational biology* **7**, e1002280 (2011).
55. Acerbi L, Ma WJ. Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search. In: *Advances in Neural Information Processing Systems* (ed^(eds) (2017).

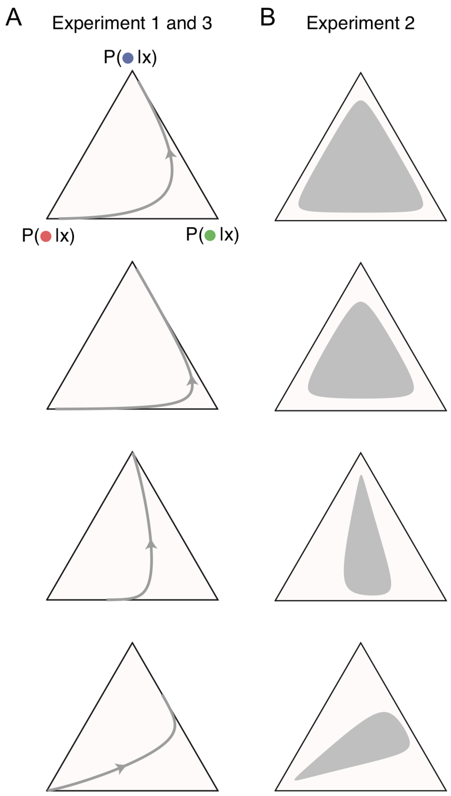
Supplementary Information

Supplementary Tables

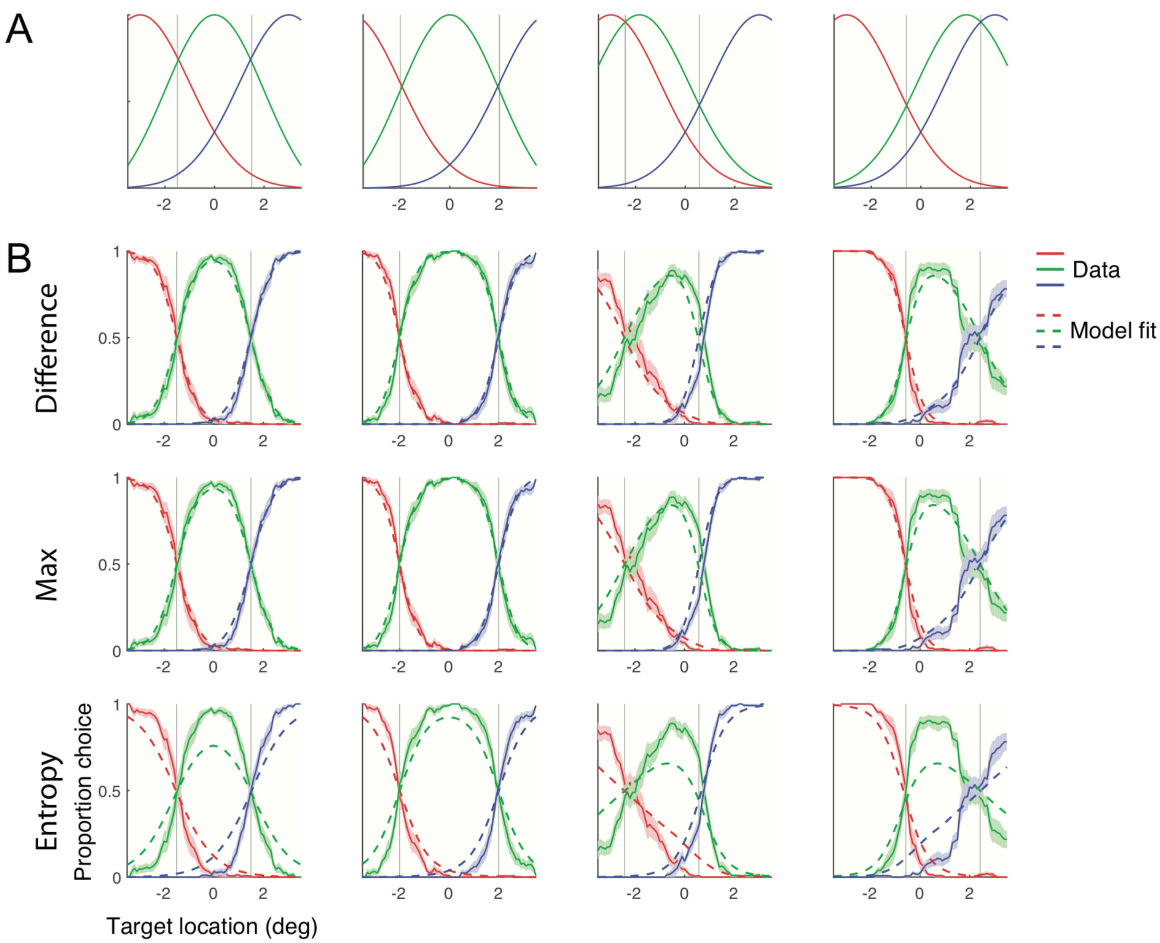
	Inference + sensory noise			Inference noise only			Sensory noise only		
	Diff	Max	Ent	Diff	Max	Ent	Diff	Max	Ent
Exp 1		27.3 (7.0)	149.4 (24)	9.9 (3.1)	34.4 (7.2)	147.7 (24)	57.3 (6.5)	98.7 (12)	317.1 (31)
Exp 2		45.9 (8.5)	151.9 (25)	13.6 (5.5)	57.8 (10)	154.8 (23)	85.5 (12)	108.5 (14)	201.1 (27)
Exp 3		10.3 (2.9)	93.2 (18)	9.27 (2.7)	17.9 (3.7)	91.5 (18)	85.5 (8.9)	139.2 (13)	327.7 (24)

Supplementary Table 1. The ΔAIC of each model, computed as the AIC of each model minus the AIC of the Difference model with both decision and sensory noise. ΔAIC is computed for individual participant. The top number in each cell is the ΔAIC averaged across participants. The numbers in the parenthesis represent one standard error of mean across participants.

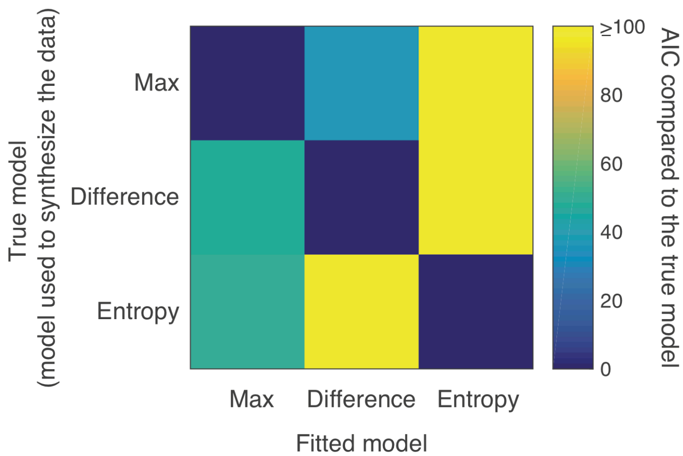
Supplementary Figures



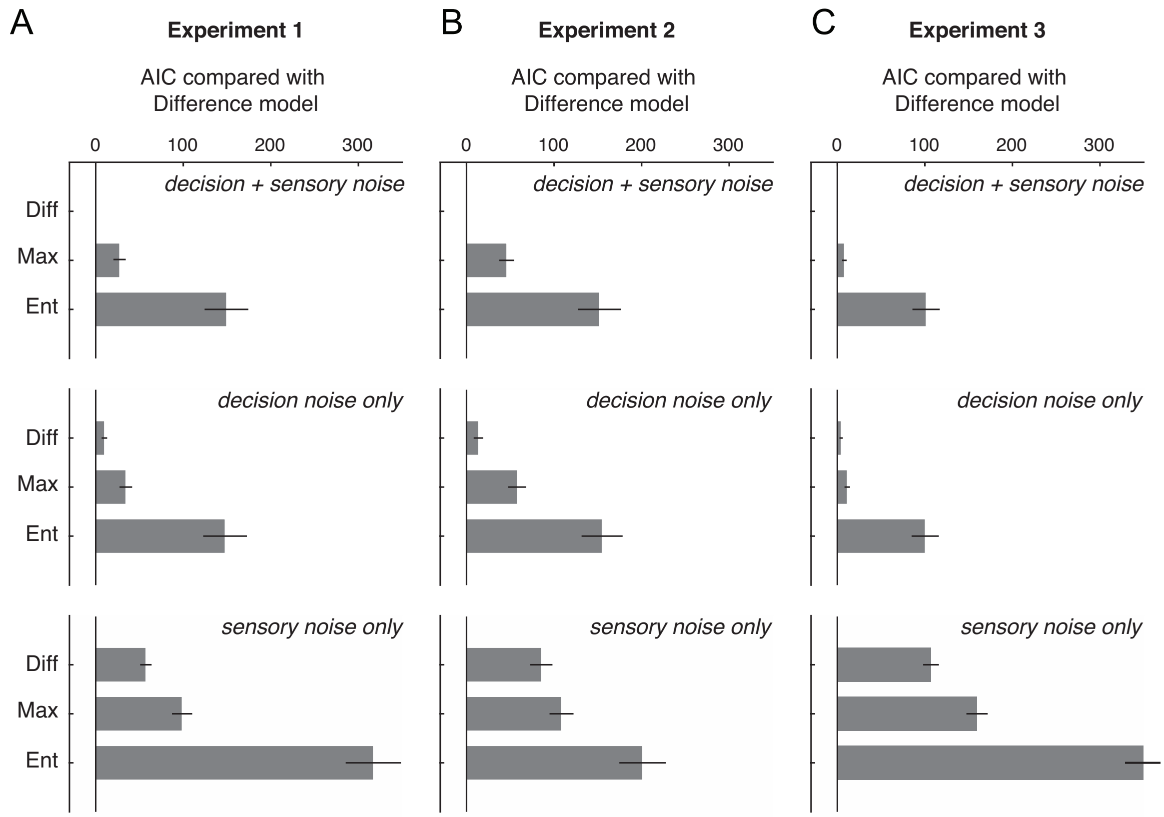
Supplementary Figure 1. Illustration of how observers' belief, posterior distribution, about the target category could change as a function of the target dot position. For illustration purpose, we considered a simplified case in which there is no sensory noise and no decision noise, so the posterior distribution only depends the target dot position and the distribution of each category. (A) Experiment 1 and 3: The four panels correspond to the four conditions depicted in Figure 1B. The gray lines and the arrows indicate the trajectory of the posterior distribution on the ternary plot as a target dot move from the left-end to the right-end of the screen. (B) Experiment 2: The four panels correspond to the four conditions depicted in Figure 1C. In the experiment, the target dot was uniformly sampled within a circle at the center of the screen with a radius of 2.6° (see Methods and S1 Experimental procedures). All possible target dot locations within the circle correspond to a range of posterior probabilities indicated by the gray region in each panel.



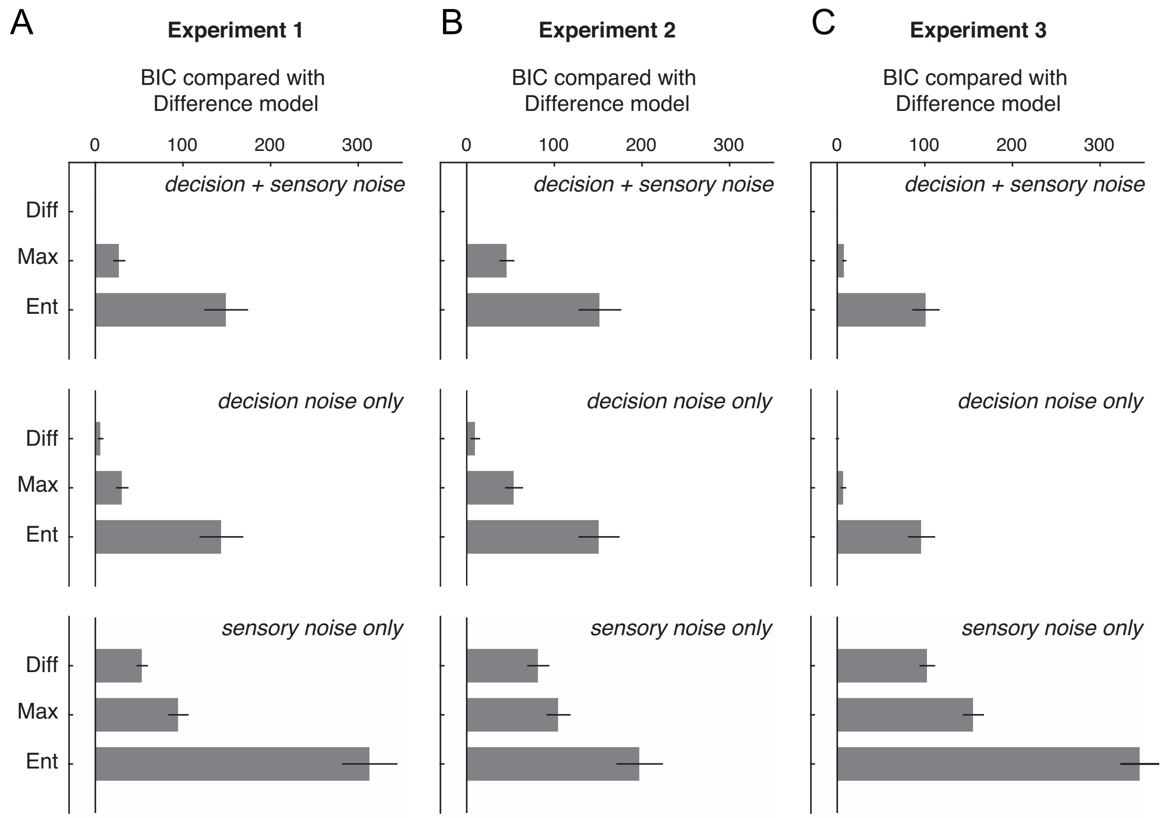
Supplementary Figure 2. Experiment 1. (A) Distribution of the reference dots in each condition. (B) The red (green, blue) lines represent the probability that the observers categorize the target dot to the red (green, blue) category as a function of the target dot location. Solid lines represent the group mean ± 1 s.e.m. The dashed lines represent the model fit averaged across individuals. In both (A) and (B), the gray vertical lines represent the boundary between two categories, the location where two categories have the same likelihood.



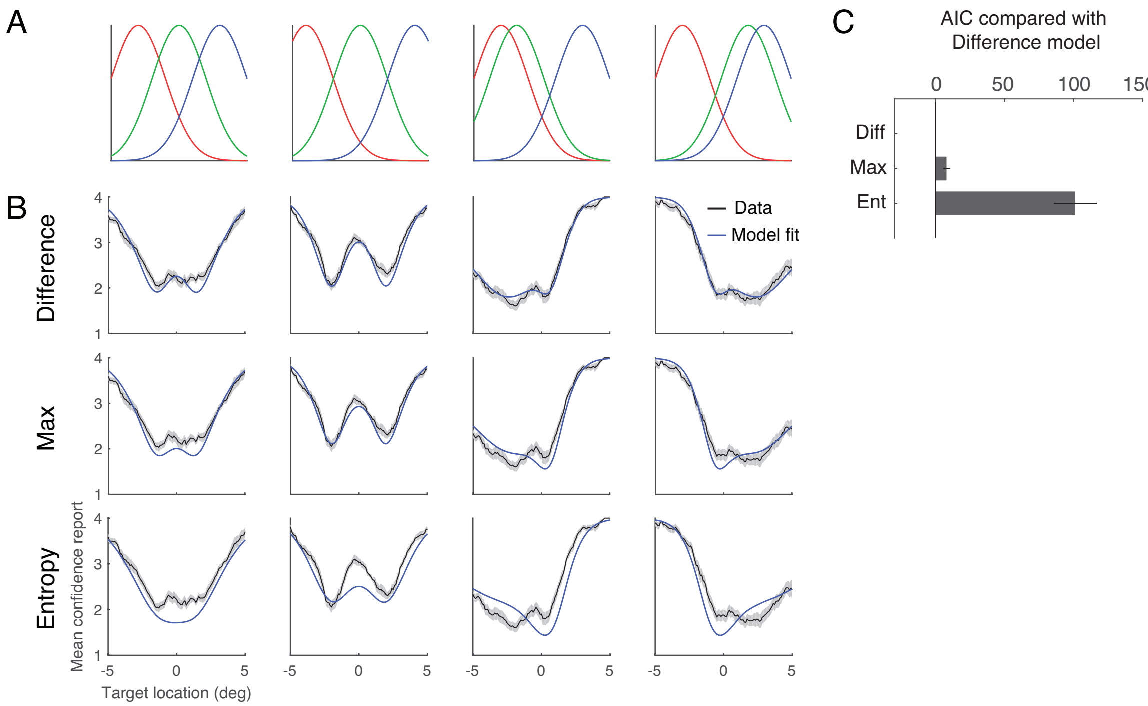
Supplementary Figure 3. Model recovery analysis. The colors represent ΔAIC of each fitted model, computed as the AIC of each fitted model minus the AIC of the fitted model using the true model.



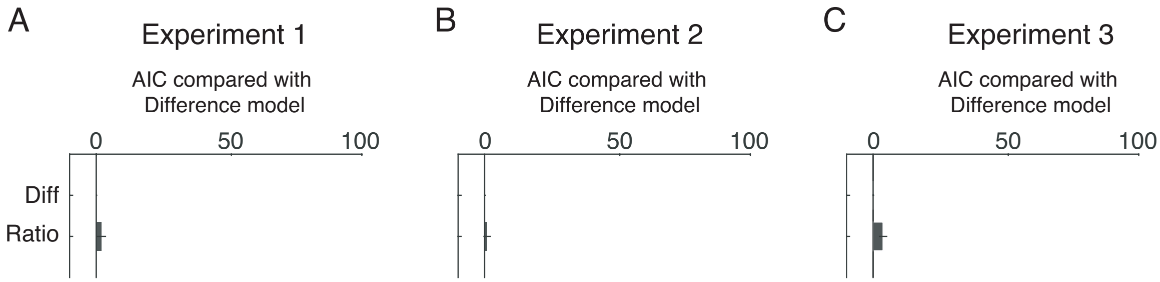
Supplementary Figure 4. Model comparison using AIC for both the full models (with both sensory and decision noise in the model; the top row) and the reduced models (with only the decision noise or only the sensory noise in the model; the middle and the bottom rows). (A) Experiment 1 (B) Experiment 2 and (C) Experiment 3. The bars represent ΔAIC (AIC of each model compared with the full Difference model) averaged across participants. The error bars represent ± 1 s.e.m across participants.



Supplementary Figure 5. Model comparison using BIC for both the full models (with both sensory and decision noise in the model; the top row) and the reduced models (with only the decision noise or only the sensory noise in the model; the middle and the bottom rows). (A) Experiment 1 (B) Experiment 2 and (C) Experiment 3. The bars represent ΔBIC (BIC of each model compared with the full Difference model) averaged across participants. The error bars represent ± 1 s.e.m across participants.



Supplementary Figure 6. Experiment 3. (A) The distribution of the reference dots in each condition. (B) Mean confidence rating as a function of target position for each of the four conditions. The black curves represent group mean ± 1 s.e.m. Blue curves represent the model fit averaged across individuals. (C) Model comparisons using ΔAIC: AIC of each model compared with the Difference model. The bars represent ΔAIC averaged across participants. The error bars represent ± 1 s.e.m across participants.



Supplementary Figure 7. Model comparison between the full Difference model and the full Ratio model using AIC (A) Experiment 1 (B) Experiment 2 and (C) Experiment 3. The bars represent ΔAIC (AIC of each model compared with the full Difference model) averaged across participants. The error bars represent ± 1 s.e.m across participants.