Supplementary Information

1 Experiment 1

1.1 Apparatus and stimuli

Apparatus. Subjects were seated in a dark room, at a viewing distance of 32 cm from the screen, with their chin in a chinrest. Stimuli were presented on a gamma-corrected 60 Hz 9.7-inch 2048-by-1536 display. The display (LG LP097QX1-SPA2) was the same as that used in the 2013 iPad Air (Apple); we chose it for its high pixel density (264 pixels/inch). The display was connected to a Windows desktop PC using the Psychophysics Toolbox extensions[46,47] for MATLAB (Mathworks).

Stimuli. The background was mid-level gray (199 cd/m^2). The stimulus was either a drifting Gabor (Subjects 3, 6, 8, 9, 10, and 11) or an ellipse (Subjects 1, 2, 4, 5, and 7). The Gabor had a peak luminance of 398 cd/m^2 at 100% contrast, a spatial frequency of 0.8 cycles per degrees of visual angle (dva), a speed of 6 cycles per second, a Gaussian envelope with a standard deviation of 4.8 dva, and a randomized starting phase. Each ellipse had a total area of 1 dva^2, and was black (0.01 cd/m^2). We varied the contrast of the Gabor and the elongation (eccentricity) of the ellipse (**Section 1.2**).

Categories. In Task A, stimulus orientations were drawn from Gaussian distributions with means $\mu_1 = -4°$ (category 1) and $\mu_2 = 4°$ (category 2) and standard deviations $\sigma_1 = \sigma_2 = 5°$. In Task B, stimulus orientations 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) (**Fig. 1b**). We chose these category means and standard deviations such that the accuracy of an optimal observer would be around 80%.

1.2 Procedure

Each subject completed 5 sessions. Each session consisted of two parts; the subject did Task A in the first part, followed by Task B in the second part, or vice versa (chosen randomly). Each part started with instruction and was followed by alternating blocks of 96 category training trials and 144 testing trials, for a total of three blocks of each type, with a block of 24 confidence training trials immediately after the first category training block. Combining all sessions and both tasks, each subject completed 4320 testing trials, 2880 category training trials, and 240 confidence training trials; we did not analyze category training or confidence training trials.

Instruction. At the start of each part of a session, subjects were shown 30 (72 in the first session) exemplar stimuli from each category. Additionally, we provided them with a printed graphic similar to **Figure 1b**, and explained how the stimuli were generated from distributions. We answered any questions.

Category training. To ensure that subjects knew the stimulus distributions well, we gave them extensive category training. Each trial proceeded as follows (**Fig. 1a**): Subjects 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 (**Fig. 1b**). Gabors had 100% contrast, and ellipses had 0.95 eccentricity (elongation). The stimulus appeared at fixation for 300 ms, replacing the fixation cross. Subjects were asked to report category 1 or category 2 by pressing a button with their left or right index finger, respectively. Subjects were able to respond immediately after the offset of the stimulus, at which point verbal correctness feedback was displayed for 1.1 s. The fixation cross then reappeared.

Confidence training. To familiarize subjects with the button mappings, they completed a short confidence training black at the start of every task. We told subjects that in this block, it would be harder to tell what the stimulus orientation was, there would be no correctness feedback, and they would be reporting their confidence on each trial in addition to their category choice. We provided them with a printed graphic similar to the buttons pictured in **Figure 1a**, indicating that they had to press one of eight buttons to indicate both 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.” Gabors had 0.4%, 0.8%, 1.7%, 3.3%, 6.7%, or 13.5% contrast, and ellipses had 0.15, 0.28, 0.41, 0.54, 0.67, or 0.8 eccentricity, chosen randomly with equal probability on each trial (**Fig. 1c**). Stimuli were only displayed for 50 ms. Trial-to-trial feedback consisted only of a message telling them which category and confidence level they had reported. Other than these changes, the trial procedure was the same as in category training.

Subjects 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.

Testing. The trial procedure in testing blocks was the same as in confidence training blocks, except that trial-to-trial feedback was completely withheld. At the end of each block, subjects were required to take at least a 30 sec break. During the break, they were shown the percentage of trials that they had correctly categorized. Subjects were also shown a list of the top 10 block scores (across all subjects, indicated by initials) for the task they had just done. This was intended to motivate subjects to score highly, and to reassure them that their scores were normal, since it is rare to score above 80% on a block.

1.3 Subjects

11 subjects (2 male), aged 20-42, participated in the experiment. Subjects received $10 per 40-60 minute session, plus a completion bonus of $15. The experiments were approved by the University Committee on Activities Involving Human Subjects of New York University. Informed consent was given by each subject before the experiment. All subjects were naïve to the purpose of the experiment. No subjects were fellow scientists.

1.4 Descriptive statistics

The following statistical differences were assessed using repeated-measures ANOVA.

In Task A, there was a significant effect of true category on category choice ($F_{1,10} = 285, P < 10^{-7}$). There was no main effect of reliability, which took 6 levels of contrast or ellipse elongation, on category choice ($F_{5,50} = 0.27, P = 0.88$). In other words, subjects were not significantly biased to respond with a particular category at low reliabilities. There was a significant interaction between reliability and true category, which is to be expected ($F_{5,50} = 59.6, P < 10^{-15}$) (**Fig. 3a**).

In Task B, there was again a significant effect of true category on category choice ($F_{1,10} = 78.3, P < 10^{-5}$). There was no main effect of reliability ($F_{5,50} = 2.93, P = 0.051$). There was again a significant interaction between reliability and true category ($F_{5,50} = 28, P < 10^{-12}$) (**Fig. 3b**).

In Task A, there was a significant effect of true category on response ($F_{1,10} = 136, P < 10^{-6}$). There was no main effect of reliability ($F_{5,50} = 0.61, P = 0.642$). There was a significant interaction between reliability and true category ($F_{5,50} = 58.7, P < 10^{-13}$) (**Fig. 3c**).

In Task B, there was a significant effect of true category on response ($F_{1,10} = 54.2, P < 10^{-6}$). There was a significant effect of reliability ($F_{5,50} = 4.84, P = 0.0128$). There was a significant interaction between reliability and true category ($F_{5,50} = 29.2, P < 10^{-8}$) (**Fig. 3d**).

In Task A, there was a main effect of confidence on the proportion of reports ($F_{3,30} = 7.75, P < 10^{-3}$); low-confidence reports were more frequent than high-confidence reports. There was no significant effect of true category ($F_{1,10} = 0.784, P = 0.397$) and no interaction between confidence and category on proportion of responses ($F_{3,30} = 1.45, P = 0.25$) (**Fig. 3e**).

In Task B, there was a main effect of confidence on the proportion of reports ($F_{3,30} = 4.36, P = 0.012$). There was no significant effect of category ($F_{1,10} = 0.22, P = 0.64$), although there was an interaction between confidence and category ($F_{3,30} = 8.37, P = 0.003$). This is likely because for task B, category 2 has a higher proportion of "easy" stimuli (**Fig. 3f**).

In both tasks, reported confidence had a significant effect on performance ($F_{3,30} = 36.9, P < 10^{-3}$). Task also had a significant effect on performance ($F_{1,10} = 20.1, P = 0.001$); although we chose the category parameters such that the performance of the optimal observer is matched, subjects were significantly better at Task A. There was no interaction between task and confidence ($F_{3,30} = 0.878, P = 0.436$) (**Fig. 3g**).

Fig. 3l-m shows psychometric choice curves for both tasks, at all 6 levels of reliability. Each point represents roughly the same number of trials.

Fig. 3n-o shows a similar set of psychometric curves. These curves differ from **Fig. 3l-m** in that they represent the mean button press rather than mean category choice.

In Task A (**Fig. 3l,n**), mean category choice and mean button press depend monotonically on orientation, with a slope that increases with reliability. In Task B (**Fig. 3m,o**), the mean category choice and mean button press tends towards category 1 when stimulus orientation is near horizontal, and tends towards category 2 when orientation is strongly tilted; this reflects the stimulus distributions.

Since some subjects only saw Gabors and some only saw ellipses, we used Spearman's rho to measure the difference in model rankings among the two groups. Spearman's rho is a measure of rank correlation for which a value of 1 indicates that the rankings of two variables are identical. Spearman's rho between Gabor and ellipse subjects for the summed LOO scores of the model groupings in **Figure 6** and **Figure S13** was 0.952 and 0.944, respectively. In both model groupings, the identities of the lowest- and highest-ranked models was the same for both Gabor and ellipse subjects. This indicates that the choice of stimulus did not have a systematic effect on model rankings.

2 Experiment 2: Separate category and confidence responses and testing feedback

This control experiment was identical to experiment 1 except for the following modifications:

- Subjects first reported choice by pressing one of two buttons with their left hand, and then reported confidence by pressing one of four buttons with their right hand.
- Subjects reported confidence in category training blocks, and received correctness feedback after reporting confidence.
- There were no confidence training blocks.
- In testing blocks, subjects received correctness feedback after each trial.
- 8 subjects (0 male), aged 19-23, participated. None had participated in experiment 1, and again, none were fellow scientists.
- Drifting Gabors were used; no subjects saw ellipses.

3 Experiment 3: Task B only

This experiment was identical to experiment 1 except for the following modifications:

- Subjects completed blocks of Task B only.
- 15 subjects (7 female), aged 19-30, participated. None had participated in experiments 1 or 2.
- Drifting Gabors were used; no subjects saw ellipses.

4 Modeling

4.1 Measurement noise

For models where the relationship between reliability (i.e., contrast or ellipse eccentricity) and noise was parametric, we assumed a power law relationship between reliability c and measurement noise s.d. σ, with additive orientation-dependent noise in the form of a rectified 2-cycle sinusoid[48]:

$$\sigma(c) = \sqrt{\gamma^2 + \alpha c^{-\beta}} + \psi|\sin 2s| \tag{1}$$

4.2 Response probability

We coded all responses as $r \in \{1, 2, \dots, 8\}$, with each value indicating category and confidence. For all models except the Linear Neural model, the probability of a single trial i is equal to the probability mass of the measurement distribution $p(x \mid s_i) = \mathcal{N}(x; s_i, \sigma_i^2)$ in a range corresponding to the subject'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. We find the boundaries $(b_{r_i-1}(\sigma_i), b_{r_i}(\sigma_i))$ in measurement space, as defined by the fitting model and parameters θ, and then compute the probability mass of the measurement distribution between the boundaries:

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

For Task A, $b_0 = -\infty°$ and $b_8 = \infty°$. For Task B, $b_0 = 0°$ 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:

$$\log p(\text{data} \mid \theta) = \sum_{i=1}^{t} \log p(r_i \mid \theta) = \sum_{i=1}^{t} \log p_\theta(r_i \mid s_i, \sigma_i) \tag{3}$$

4.3 Model specification

4.3.1 Bayesian

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

$$d = \log\frac{p(C=1 \mid x)}{p(C=2 \mid x)} = \log\frac{p(x \mid C=1)}{p(x \mid C=2)} + \log\frac{p(C=1)}{p(C=2)} \tag{4}$$

To get d_A and d_B, we need to find the task-specific expressions for $p(x \mid C)$. 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:

$$p(x \mid C) = \int p(x \mid s)p(s \mid C)ds \tag{5}$$

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

$$p(x \mid C) = \int \mathcal{N}(s; x, \sigma^2)\mathcal{N}(s; \mu_C, \sigma_C^2)ds = \mathcal{N}(x; \mu_C, \sigma^2 + \sigma_C^2) \tag{6}$$

Plugging in the task- and category-specific μ_C and σ_C, and substituting these expressions back into equation (4), we get:

$$d_\mathrm{A} = \frac{2x\mu_1}{\sigma^2+\sigma_1^2} + \log\frac{p(C=1)}{p(C=2)} \tag{7}$$

$$d_\mathrm{B} = \frac{1}{2}\log\frac{\sigma^2+\sigma_2^2}{\sigma^2+\sigma_1^2} - \frac{\sigma^2-\sigma_1^2}{2(\sigma^2+\sigma_1^2)(\sigma^2+\sigma_2^2)}x^2 + \log\frac{p(C=1)}{p(C=2)} \tag{8}$$

In all of our Bayesian models, 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 (and not confidence). k_0 is fixed at $-\infty$ and k_8 is fixed at ∞. In all models, the observer chooses category 1 when d is positive.

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

Levels of strength.

In Bayes$_\mathrm{Ultrastrong}$, $\mathbf{k}$ is symmetric across k_4: $k_{4+j} - k_4 = k_4 - k_{4-j}$ for $j \in \{1,2,3\}$. Furthermore, in Bayes$_\mathrm{Ultrastrong}$, $\mathbf{k}_\mathrm{A} = \mathbf{k}_\mathrm{B}$. So Bayes$_\mathrm{Ultrastrong}$ has a total of 4 free boundary parameters: k_1, k_2, k_3, k_4. Bayes$_\mathrm{Ultrastrong}$ is equivalent to the observer comparing $|d_\mathrm{A}|$ and $|d_\mathrm{B}|$ to a task-general set of boundaries to determine the response (**Fig. S1**, left column).

Bayes$_\mathrm{Strong}$ is identical to Bayes$_\mathrm{Ultrastrong}$ except that $\mathbf{k_A}$ is allowed to differ from $\mathbf{k_B}$. So Bayes$_\mathrm{Strong}$ has a total of 8 free boundary parameters: $k_{1\mathrm{A}}, k_{2\mathrm{A}}, k_{3\mathrm{A}}, k_{4\mathrm{A}}, k_{1\mathrm{B}}, k_{2\mathrm{B}}, k_{3\mathrm{B}}, k_{4\mathrm{B}}$. Bayes$_\mathrm{Strong}$ is equivalent to the observer comparing $|d_\mathrm{A}|$ to one set of boundaries, and $|d_\mathrm{B}|$ to a different set of boundaries (**Fig. S1**, middle

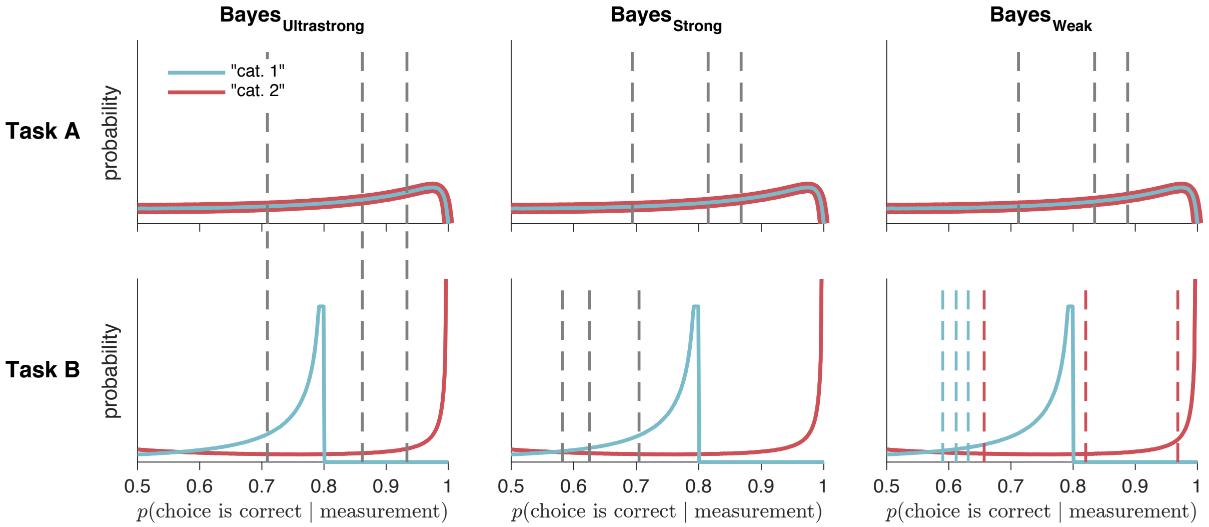

Figure S1: Distributions of posterior probabilities of being correct, with confidence criteria for Bayesian models with three different levels of strength. Solid lines represent the distributions of posterior probabilities for each category and task in the absence of measurement noise. Dashed lines represent confidence criteria, generated from the mean of subject 4's posterior distribution over parameters. Each model has a different number of sets of mappings between posterior probability and confidence report. In $\text{Bayes}_{\text{Ultrastrong}}$, there is one set of mappings. In $\text{Bayes}_{\text{Strong}}$, there is one set for Task A, and another for Task B. In $\text{Bayes}_{\text{Weak}}$, there is one set for Task A, and one set for each reported category in Task B. Plots were generated from the mean of subject 4's posterior distribution over parameters as in **Figure 2**.

column).

$\text{Bayes}_{\text{Weak}}$ is identical to $\text{Bayes}_{\text{Strong}}$ except that symmetry is not enforced for $\mathbf{k_B}$. So $\text{Bayes}_{\text{Weak}}$ has a total of 11 free boundary parameters: $k_{1\text{A}}, k_{2\text{A}}, k_{3\text{A}}, k_{4\text{A}}, k_{1\text{B}}, k_{2\text{B}}, k_{3\text{B}}, k_{4\text{B}}, k_{5\text{B}}, k_{6\text{B}}, k_{7\text{B}}$. $\text{Bayes}_{\text{Weak}}$ is equivalent to the observer comparing $|d_\text{A}|$ to a set of boundaries, and d_B to a different set of boundaries (**Fig. S1**, right column).

Decision boundaries.

In the Bayesian models without d noise, we plug the parameters $\mathbf{k}$ into the left-hand side of equations (7) and (8) and solve for x at the fitted levels of σ. These values were used as the measurement boundaries $\mathbf{b}(\sigma)$.

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 equations (7) and (8) 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[49]. We then computed 101 response probabilities for each trial (**Section 4.2**), one for each draw of η_d, and computed the weighted average according to $p(\eta_d)$.

4.3.2 Fixed

In Fixed, 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 $b_r(\sigma) = k_r$.

4.3.3 Lin and Quad

In Lin and Quad, 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 $b_r(\sigma) = k_r + m_r\sigma$ (Lin) or $b_r(\sigma) = k_r + m_r\sigma^2$ (Quad).

4.3.4 Orientation Estimation

In Orientation Estimation, the observer uses their knowledge of the stimulus distributions to compute a maximum a posteriori estimate of the stimulus:

$$
\begin{aligned}
\hat{s} &= \underset{s}{\operatorname{argmax}}\; p(s \mid x) && (9)\\
&= \underset{s}{\operatorname{argmax}}\; p(x \mid s)p(s) && (10)\\
&= \underset{s}{\operatorname{argmax}}\; \left[\mathcal{N}(s; x, \sigma^2)(p(s \mid C=1) + p(s \mid C=2))\right] && (11)
\end{aligned}
$$

The observer then compares $\hat{s}$ to a set of boundaries $\mathbf{k}$ to determine category and confidence response.

Decision boundaries. To find the decision boundaries in measurement space, we used *gmm1max_n2_fast* from Luigi Acerbi's gmm1 Gaussian mixture model toolbox to solve equation (11), computing a lookup table containing the value of $\hat{s}$ as a function of x and σ[35]. We then found, using linear interpolation, the values of x corresponding to σ and the free parameters $\mathbf{k}$. These values were used as the measurement boundaries $\mathbf{b}(\sigma)$.

4.3.5 Linear Neural

In this section, $\mathbf{r}$ refers to neural activity, not button responses.

This model is different from all other models in that the generative model does not include measurement x. The model can be derived as follows.

All neurons have Gaussian tuning curves with variance σ^2_{TC} and gain $g = \frac{1}{\sigma^2}$. Tuning curve means are contained in the vector of preferred stimuli $\tilde{\mathbf{s}}$. The number of spikes in the population is $\mathbf{r} \sim \text{Poisson}(g\mathcal{N}(s; \tilde{\mathbf{s}}, \sigma^2_{\text{TC}}))$. Neural weights are a linear function of the preferred stimuli: $\mathbf{w} = a\tilde{\mathbf{s}}$.

On each trial, we get some quantity that is a weighted sum of each neuron's activity, $z = \mathbf{w} \cdot \mathbf{r}$. $\mathbb{E}\left[z \mid s\right] = \mathbf{w} \cdot \mathbb{E}\left[\mathbf{r} \mid s\right] = ag\sum_j \tilde{s}_j \exp\left(-\frac{(s-\tilde{s}_j)^2}{2\sigma^2_{\text{TC}}}\right)$.

Rather than sum over all neurons, we assume an infinite number of neurons uniformly spanning all possible preferred stimuli $\tilde{s}$. This allows us to replace the sum with an integral. The expected value of z is $ag\int \tilde{s}\exp\left(-\frac{(s-\tilde{s}_j)^2}{2\sigma^2_{\text{TC}}}\right)d\tilde{s} = ags\sqrt{2\pi\sigma^2_{\text{TC}}}$. The variance of z is $\sum_j w_j^2 f_j(s) = ag\int \tilde{s}^2\exp\left(-\frac{(s-\tilde{s})^2}{2\sigma^2_{\text{TC}}}\right)d\tilde{s} = ag\sqrt{2\pi\sigma^2_{\text{TC}}}(\sigma^2_{\text{TC}} + s^2)$.

Now that we have the mean and variance of z, we are going to assume that z is normally distributed. This is equivalent to assuming that there are a high number of spikes, because the Poisson distribution approximates the normal distribution as the rate parameter becomes high. To compute response probability, we fit neural activity boundaries $\mathbf{k}$, and replace equation (2) with:

$$p_\theta(r_i \mid s_i, \sigma_i) = \int_{k_{r_i-1}}^{k_{r_i}} \mathcal{N}(z; ags_i\sqrt{2\pi\sigma^2_{\mathrm{TC}}}, ag\sqrt{2\pi\sigma^2_{\mathrm{TC}}}(\sigma^2_{\mathrm{TC}} + s_i^2))dz \tag{12}$$

4.4 Lapse rates

In confidence and category 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}.

4.5 Parameterization

Because of tradeoffs when directly fitting parameters γ, α, β, we re-parameterized equation (1) as

$$\sigma(c) = \sqrt{\sigma_{\mathrm{L}}^2 + \frac{(\sigma_{\mathrm{L}}^2 - \sigma_{\mathrm{H}}^2)(c^{-\beta} - c_{\mathrm{L}}^{-\beta})}{(c_{\mathrm{L}}^{-\beta} - c_{\mathrm{H}}^{-\beta})}} + \psi|\sin 2s|, \tag{13}$$

where c_{L} and c_{H} were the values of the lowest and highest reliabilities used. This way, σ_{L} and σ_{H} were free parameters that determined the s.d. of the measurement distributions for the lowest and highest reliabilities, with β determining the curvature of the function between the two reliabilities. For models where the relationship between reliability and noise was non-parametric, the first term in equation (1) was replaced with free s.d. parameters ($\sigma_{\text{rel. }1}, \ldots, \sigma_{\text{rel. }6}$) corresponding to each of the six reliability levels.

For models where subjects had incorrect knowledge about their measurement noise, we fitted two sets of uncertainty-related parameters. One set was for the generative noise (used in equation (2)), and the other set was for the subject’s believed noise (used in equations (7), (8), and (11)).

All parameters that defined the width of a distribution ($\sigma_{\mathrm{L}}, \sigma_{\mathrm{H}}, \sigma_d, \sigma_1, \ldots$) were sampled in log-space and exponentiated during the computation of the log likelihood. See *model_parameters.xls* for a complete list of each model’s parameters.

4.6 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 for both model comparison and for plotting model fits. To sample from the posterior, we use an expression for the unnormalized log posterior:

$$\log p(\theta \mid \text{data}) = \log p(\text{data} \mid \theta) + \log p(\theta) + \text{constant}, \tag{14}$$

where we assumed a factorized prior over each parameter j:

$$\log p(\theta) = \sum_{j=1}^{n} \log p(\theta_j), \tag{15}$$

where j is the parameter index and n is the number of parameters. We used log-normal priors for parameters that were standard deviations. For the other parameters, we took uniform priors over reasonable, sufficiently large ranges[35].

We sampled from the probability distribution using a Markov Chain Monte Carlo (MCMC) method, slice sampling[50]. For each model and dataset combination, we ran between 4 and 7 parallel chains with random starting points. For each chain, we took 40,000 to 600,000 total samples (depending on model computational time), discarded the first third of the samples, and kept 10,000 of the remaining samples, randomly selected. All samples with log posteriors less than 40 below the log posterior of the sample with the highest log posterior were discarded. Marginal probability distributions of the sample log likelihoods were visually checked for convergence across chains. In total we had 842 model and dataset combinations, with a median of 26,668 kept samples (IQR = 13,334).

4.7 Model comparison

A more complex model is likely to fit a dataset better than a simpler model, even if only by chance. Since we are interested in our models' predictive accuracy for unobserved data, it is important to choose a metric for model comparison that takes the complexity of the model into account, avoiding the problem of overfitting. Roughly speaking, there are two ways to compare models: information criteria and cross-validation.

Information criteria such as AIC, BIC, AICc, and WAIC, are widely used metrics that add corrections to the negative log likelihood of the dataset. Most information criteria are based on a point estimate for θ, typically θ_{MLE}, the θ that maximizes the log likelihood of the dataset (equation (3)). AIC adds a correction for the number of parameters n to the log likelihood of the dataset: $\text{AIC} = -2\sum_{i=1}^{t} \log p(r_i \mid \theta_{\text{MLE}}) + 2n$. BIC and AICc add corrections that are functions of the number of parameters and the number of trials. WAIC is a more Bayesian approach which is based on samples from the full posterior of θ (equation (14)), and adds a correction for the effective number of parameters[51]. Although information criteria are computationally convenient, they are based on asymptotic results and assumptions about the data that may not always hold[51].

An alternative way to estimate predictive accuracy for unobserved data is to cross-validate, fitting the model to training data and evaluating the fit on held out data. Leave-one-out cross-validation (LOO) is the most thorough way to cross-validate, but is very computationally intensive; it requires that you fit your model t times, where t is the number of trials. Here we use a method proposed by Vehtari *et al.*[32] for approximating LOO with the full posterior obtained through MCMC:

$$\text{LOO} = \sum_{i=1}^{t} \log \frac{\sum_u w_{i,u} p(r_i \mid \theta_u)}{\sum_u w_{i,u}}, \tag{16}$$

where $w_{i,u}$ is the importance weight of trial i for sample u, and θ_u is the u-th sampled set of parameters.

In model comparison studies, we are often concerned about whether our conclusions are dependent on the choice of model comparison metric. Would our model rankings be substantially different if we had used an information criterion instead of LOO?

We computed AIC, BIC, AICc, WAIC, and LOO for all models in the 8 model groupings in **Figure S13-Figure S20**, multiplying the information criteria by $-\frac{1}{2}$ to match the scale of LOO. For AIC, BIC, and AICc, we used the parameter sample with the highest log likelihood as our estimate of θ_{MLE}. Then we computed Spearman's rho for every possible pairwise comparison of model comparison metrics for all model and dataset combinations, producing 80 total values (8 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.

4.8 Visualization of model fits

Model fits were plotted by bootstrapping synthetic group datasets with the following procedure: For each task, model, and subject, 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 subject. We randomly selected a number of synthetic datasets equal to the number of subjects to create a synthetic group dataset. For each synthetic group dataset, we computed the mean output (e.g., button press, confidence, performance) per 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 ±1 s.e.m. of synthetic group datasets.

For plots with orientation on the horizontal axis (e.g., **Figure 3j-o**), stimulus orientation was binned according to quantiles of the task-dependent stimulus distributions so that each point consisted of roughly the same number of trials. For each task, 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, adjusting the horizontal axis tick marks appropriately.

4.9 Model recovery

We performed a model recovery analysis[52] to test our ability to distinguish our core models. We generated synthetic datasets from each of the 8 core models in **Figure 6**, for both Tasks A and B, using the same sets of stimuli that were originally randomly generated for each of the 11 subjects. To ensure that the statistics of the generated responses were similar to those of the subjects, we generated responses to these stimuli from 4 of the randomly chosen parameter estimates obtained via MCMC sampling (as described in **Section 4.6**) for each subject and model. In total, we generated 352 datasets (8 generating models × 11 subjects × 4 datasets). We then fit all 8 models to every dataset. For computational efficiency, we fit the models with maximum likelihood estimation of parameters by an interior-point constrained optimization (MATLAB's *fmincon*) rather than with the more time-consuming MCMC sampling used for subject data. We then computed AIC scores from the resulting fits.

We found that the true generating model was the best-fitting model, on average, in all cases (**Fig. S2**). Overall, AIC "selected" the correct model (i.e., AIC scores were lowest for the model that generated the data) for 86.6% of the datasets, indicating that our models are distinguishable. We believe that these results would hold if we used the same fitting and comparison procedures as for the subject data (MCMC and LOO, rather than MLE and AIC), because we found that AIC and LOO scores gave us near-identical model rankings for data from real subjects (**Section 4.7**).

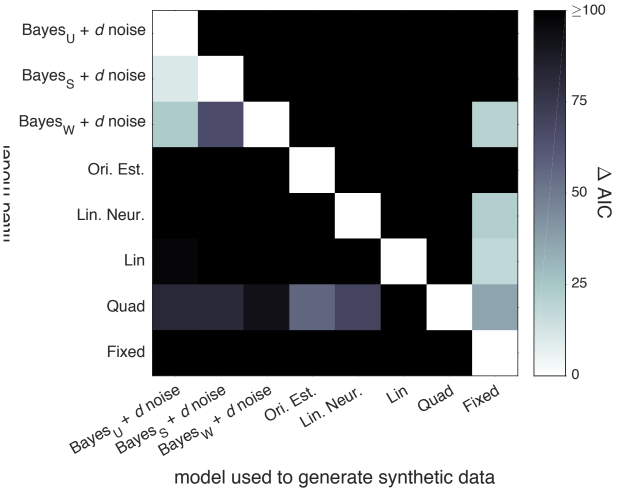

Figure S2: Model recovery analysis. Shade represents the difference between the mean AIC score (across datasets) for each fitted model and for the one with the lowest mean AIC score. White squares indicate the model that had the lowest mean AIC score when fitted to data generated from each model. The squares on the diagonal indicate that the true generating model was the best-fitting model, on average, in all cases.

5 Neural networks

5.1 Architecture

In this section, r and $\mathbf{r}$ refer to neural activity, not button responses.

We trained 3-layer feedforward neural networks to perform Task B[42]. The architecture, described below, is pictured in **Figure 7a**. The input units were 50 independent Poisson neurons. The mean number of spikes per trial was determined by Gaussian tuning curves with baselines, such that the neurons had spike count

$$\mathbf{r}_{\text{input}} \sim \text{Poisson}\left(g\mathcal{N}(s, \tilde{\mathbf{s}}, \sigma^2_{\text{TC}}) + \zeta\right). \tag{17}$$

$\tilde{\mathbf{s}}$ is the vector of preferred stimuli, which were linearly spaced from −40° to 40°. All neurons had tuning curve width $\sigma^2_{\text{TC}} = 100$ and baseline $\zeta = 0.025$ (to limit the number of trials with zero spikes). Gain g varied from trial-to-trial, and was derived from fits to subject data. Input units were connected, all-to-all, to 200 hidden rectified linear units with responses

$$\mathbf{r}_{\text{hidden}} = \max(0, \mathbf{W}_{\text{input}}\mathbf{r}_{\text{input}}), \tag{18}$$

where $\mathbf{W}_{\text{input}}$ was the weight matrix applied to the input units. Both input and hidden layers included a bias unit with a constant response of 1, which, when multiplied by the fitted weights, effectively adds a fitted bias to the hidden units and output unit. Hidden units were connected to a sigmoidal output unit with response

$$r_{\text{output}} = \frac{1}{1 + \exp(-\mathbf{w}_{\text{hidden}} \cdot \mathbf{r}_{\text{hidden}})}, \tag{19}$$

where $\mathbf{w}_{\text{hidden}}$ was the weight vector applied to the hidden units.

5.2 Training networks and generating datasets

Stimuli s were drawn from the same distributions used for the human experiments in Task B. Gains differed from trial-to-trial; to ensure that the reliability of the information available to the networks was similar to the reliability of the subjects' sensory information, the gains were derived from the fits to the Task A choice data in experiment 1. Indeed, the performance of the networks roughly matched the performance of the subjects (**Fig. S3c**). We used 15 different values for the number of training trials, ranging from 10 to 4.6×10^5, logarithmically spaced (**Figure 7b-d** only depicts results from the most highly trained networks).

We used standard back-propagation[54] to minimize cross-entropy between network output and category labels; as with the human subjects, the networks did not receive probabilistic feedback during training. Weights were initialized to small random values drawn from a zero-mean Gaussian distribution with s.d. 0.05. We used mini-batch gradient descent with a batch size of 10, over a single epoch. We used L2 regularization with regularization term $\alpha = 10^{-4}$.

We decoded the optimal posterior $p(C = 1 \mid \mathbf{r}_{\text{input}})$, which allowed us to compute fractional information loss. Fractional information loss was defined as the KL-divergence between $p(C = 1 \mid \mathbf{r}_{\text{input}})$ and r_{output}, normalized by the mutual information between the category labels and $\mathbf{r}_{\text{input}}$[55].

Learning rate η decreased as a function of the batch number j:

$$\eta_j = \frac{\eta_0}{1 + \tau j} \tag{20}$$

We used a constrained pattern search optimization (MATLAB's *patternsearch*) to find, for the gains associated with each subject, the η_0 and τ that minimized fractional information loss on a validation set. We used *patternsearch* because, unlike *fmincon*, it is well-suited for optimizing stochastic objective functions.

From each trained network, we generated a test set consisting of 2160 trials, the same number of Task B trials completed by subjects in experiment 1. r_{output} was mapped onto the 8 category and confidence responses using quantiles. We produced datasets from 4 separately trained networks for gains associated with each subject, generating 660 datasets in total (15 numbers of training trials × 11 subject-derived sets of gains × 4 datasets).

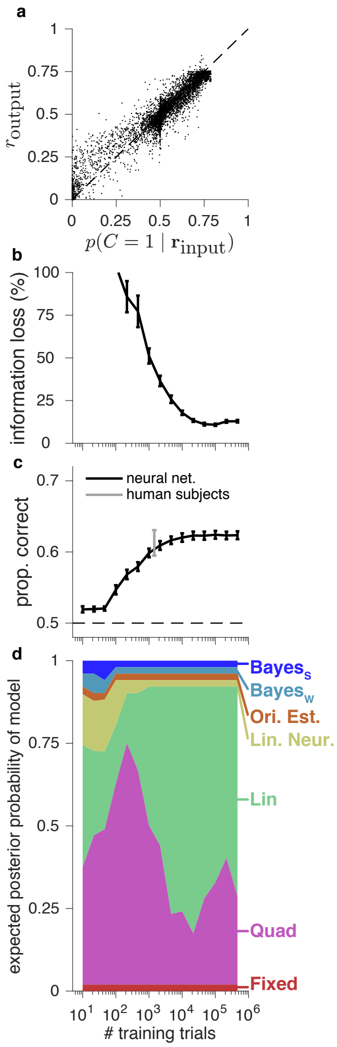

Figure S3: Neural network optimality, task performance, and model comparison. (**a**) Posterior probabilities decoded optimally from input unit activity, scattered against network output. Each point is a test trial from a neural network that was trained on the maximum number of training trials. For clarity, a randomly selected subset of test trials is plotted. (**b**) Fractional information loss as a function of the number of training trials. Error bars represent ±1 s.e.m. across the means of datasets generated with the gains derived from each subject. (**c**) Black line indicates network test performance as a function of the number of training trials. Gray error bar indicates ±1 s.e.m. for the Task B performance of subjects in experiment 1; all subjects completed 1440 training trials. (**d**) Expected posterior probabilities of each model[53] as a function of the number of training trials.

We found that r_{output} was a fairly good approximation of the optimal posterior $p(C = 1 \mid \mathbf{r}_{\text{input}})$, with some positive bias when the posterior was low (**Fig. S3a**). We also found that information loss and performance went down as the number of training trials increased (**Fig. S3b-c**). Both information loss and performance appear to reach asymptote around 2×10^4 trials; therefore, it is unlikely that our results are due to insufficient training.

5.3 Behavioral models

We fit the 660 network-generated datasets, obtaining AIC scores for each dataset and model. The fitted models were identical to the 7 Task B models described in the text (and shown in **Figure S15** and **Table S3**), except that we removed

the following mechanisms that we knew not to be present in the neural network generative process:

- All lapse rates except for a uniform lapse rate over all 8 responses
- Orientation-dependent noise
- d noise (applicable to Bayesian models only)

As with model recovery (**Section 4.9**), we used MLE and AIC (rather than MCMC and LOO) for computational efficiency, due to the large number of datasets being fitted.

After fitting, we used the *spm_BMS* function in the SPM12 software package to compute the expected posterior probability distribution over models at each number of training trials, using a random effects method of Bayesian model selection[53]. We found that Lin and Quad fit the data best, with Quad fitting best for data generated from less well-trained networks, and Lin fitting best for data generated from highly trained networks (**Fig. S3d**). This transition is consistent with previous results[42]. We plotted the summed AIC differences for data generated from the most highly trained networks in **Figure S4**, blue bars.

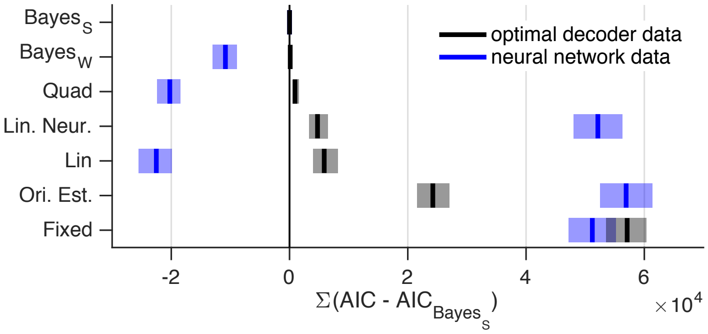

Figure S4: Model comparison for data generated from an optimal decoder of spikes, and from neural networks trained with 4.6×10^5 training trials. Models are ordered by quality of fit to optimal decoder data. As in **Figure 7b**, but for all models.

5.4 Control

To ensure that our results were not due to overfitting, we performed a model recovery analysis. This was similar to the previously described model recovery, except that we used input layer activity $\mathbf{r}_{\text{input}}$ rather than x and σ. For the $\mathbf{r}_{\text{input}}$ used in the test set of the 44 most highly trained networks (11 subject-derived sets of gains × 4 datasets), we decoded the per-trial optimal posteriors $p(C = 1 \mid \mathbf{r}_{\text{input}})$. We then mapped the posterior onto the 8 category and confidence responses using quantiles (this is also how we mapped the trained networks' r_{output} onto responses, see above). We then fit these datasets with the same models used to fit the datasets produced by the trained networks. We found that $\text{Bayes}_{\text{Strong}}$ was the best-fitting model (**Fig. S4**, black). This suggests that the architecture or the training procedure of the neural networks constrains the type of behavior that can be produced.

6 Data and code availability

All data and code used for running experiments, model fitting, and plotting is available on GitHub.

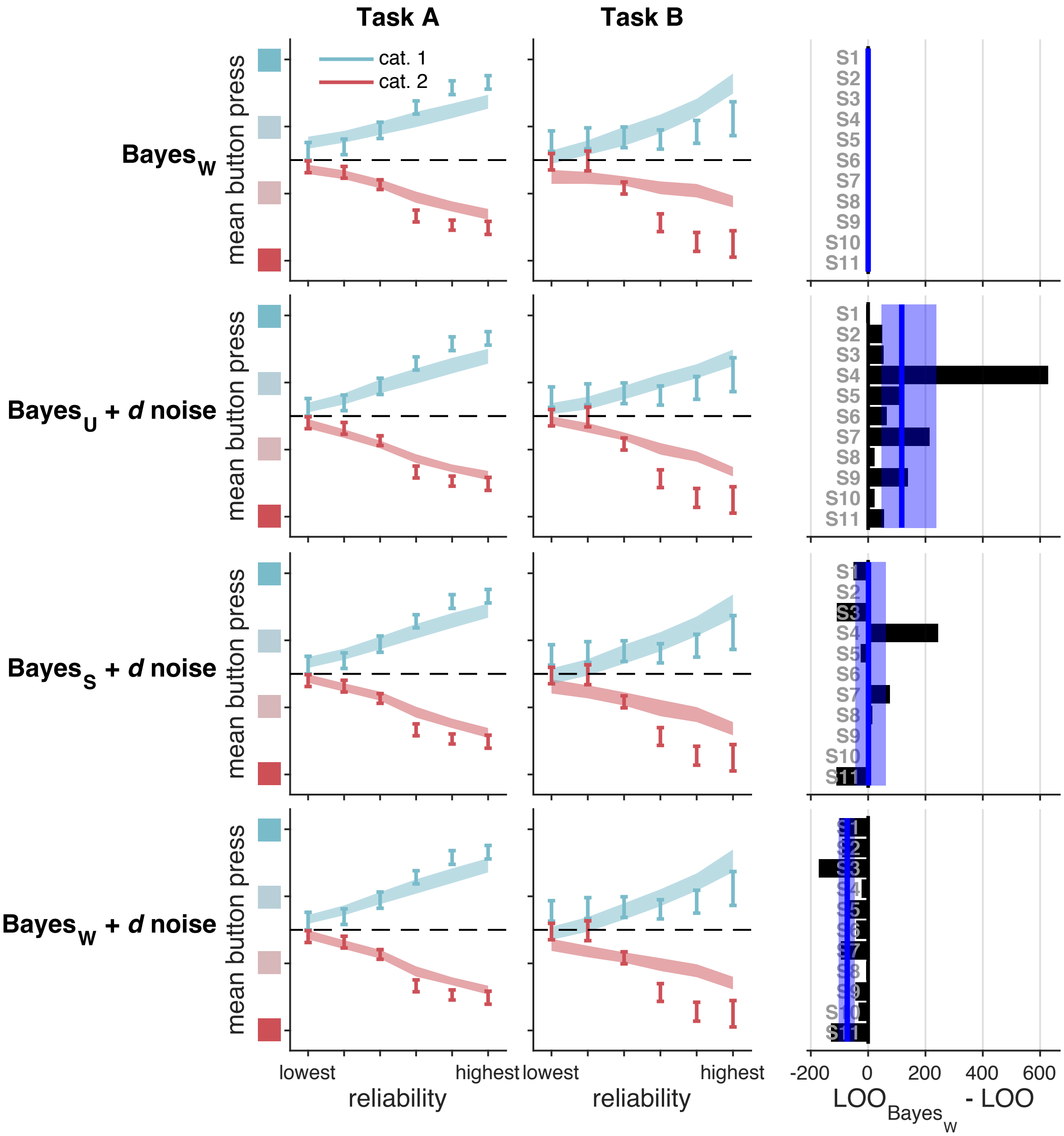

Figure S5: Model fits and model comparison for best-fitting Bayesian model without d noise, and for all Bayesian models with d noise, as in **Figure 4**.

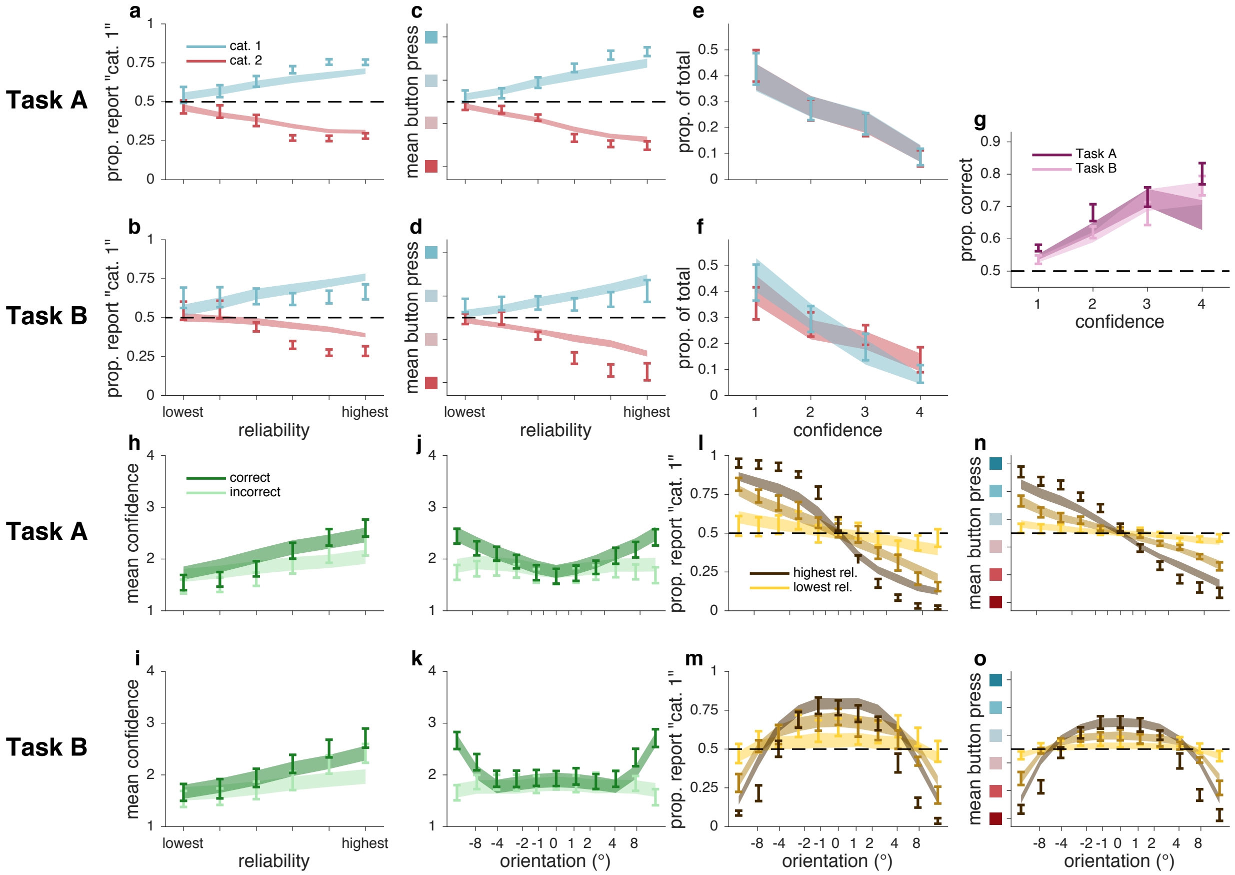

Figure S6: Bayes$_{\text{Ultrastrong}}$ + d noise fits, as in **Figure 3**.

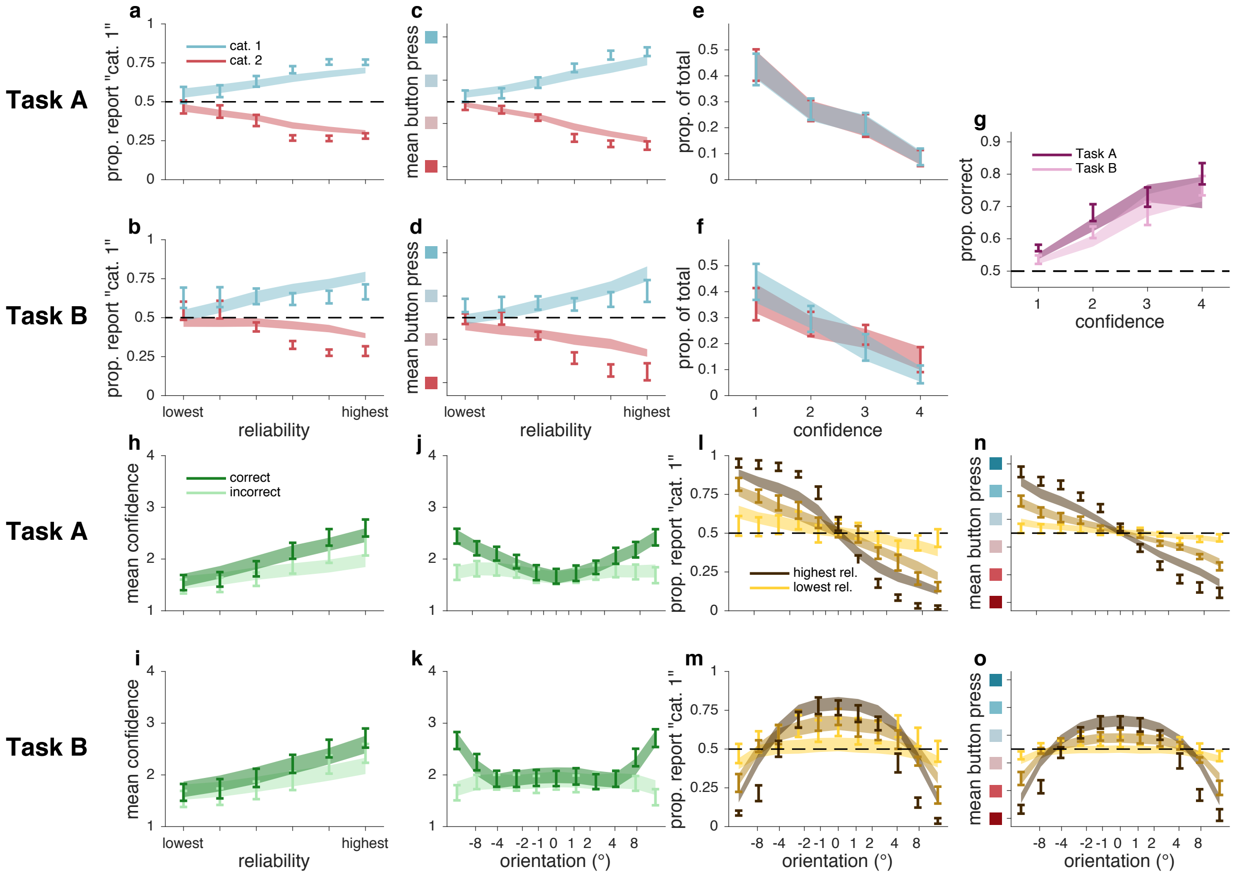

Figure S7: Bayes$_{\text{Strong}}$ + d noise fits, as in **Figure 3**.

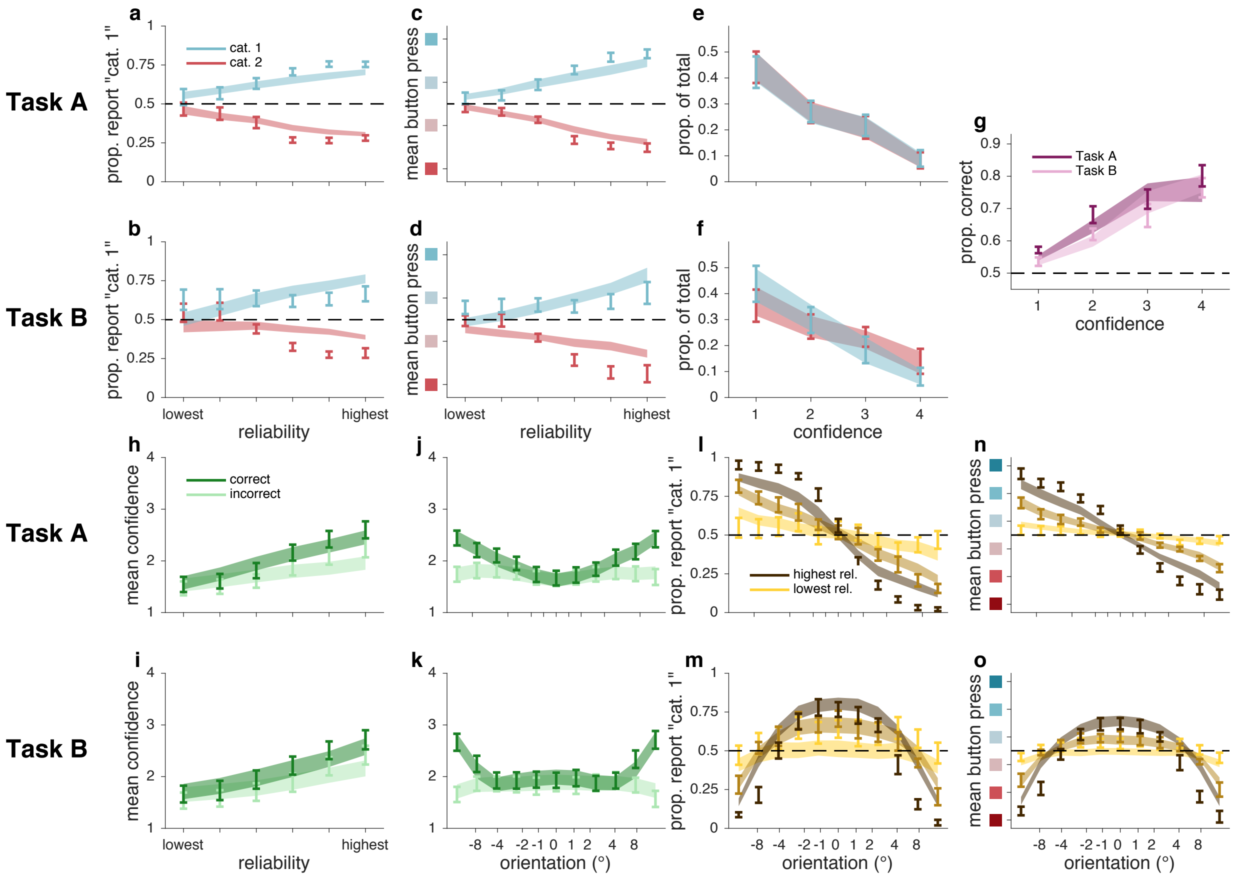

Figure S8: $\text{Bayes}_{\text{Weak}} + d$ noise fits, as in **Figure 3**.

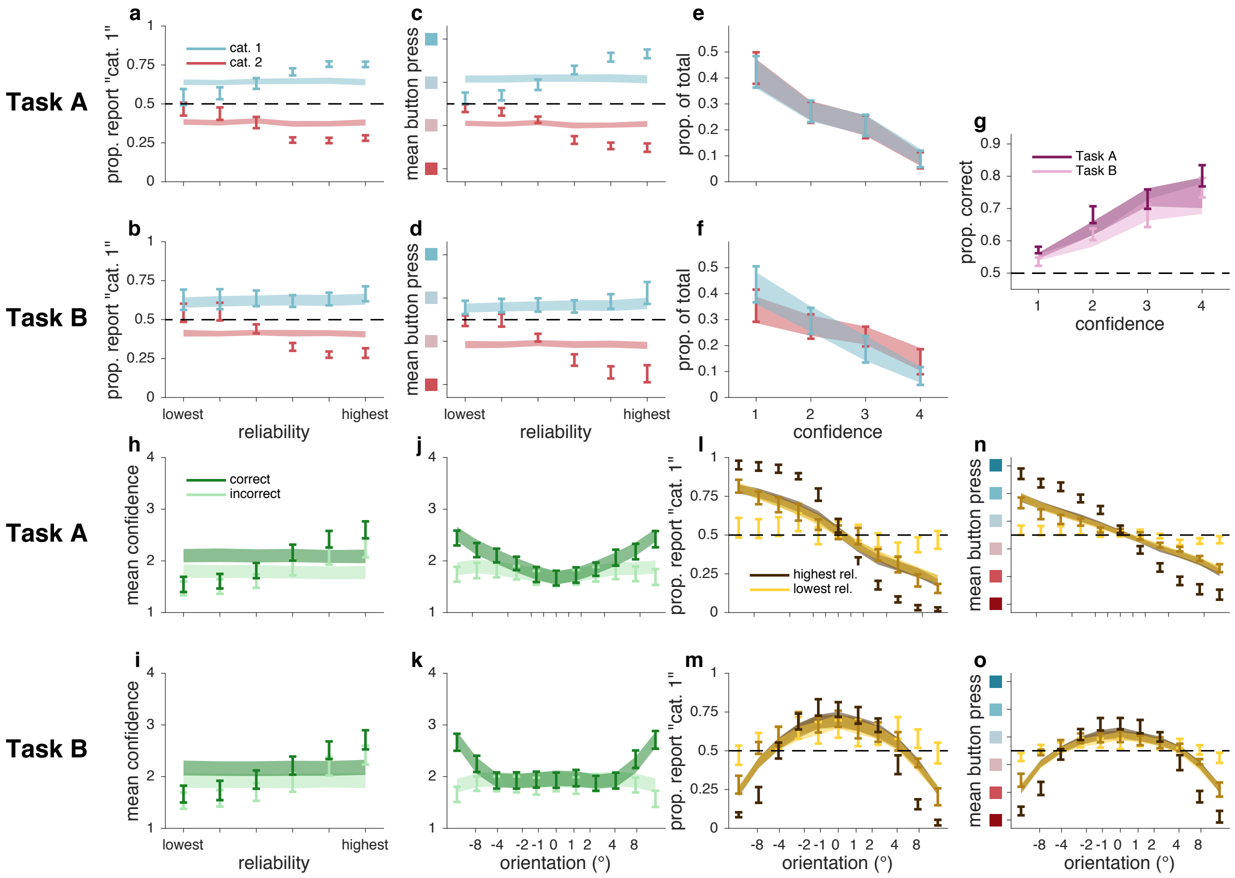

Figure S9: Fixed fits, as in **Figure 3**.

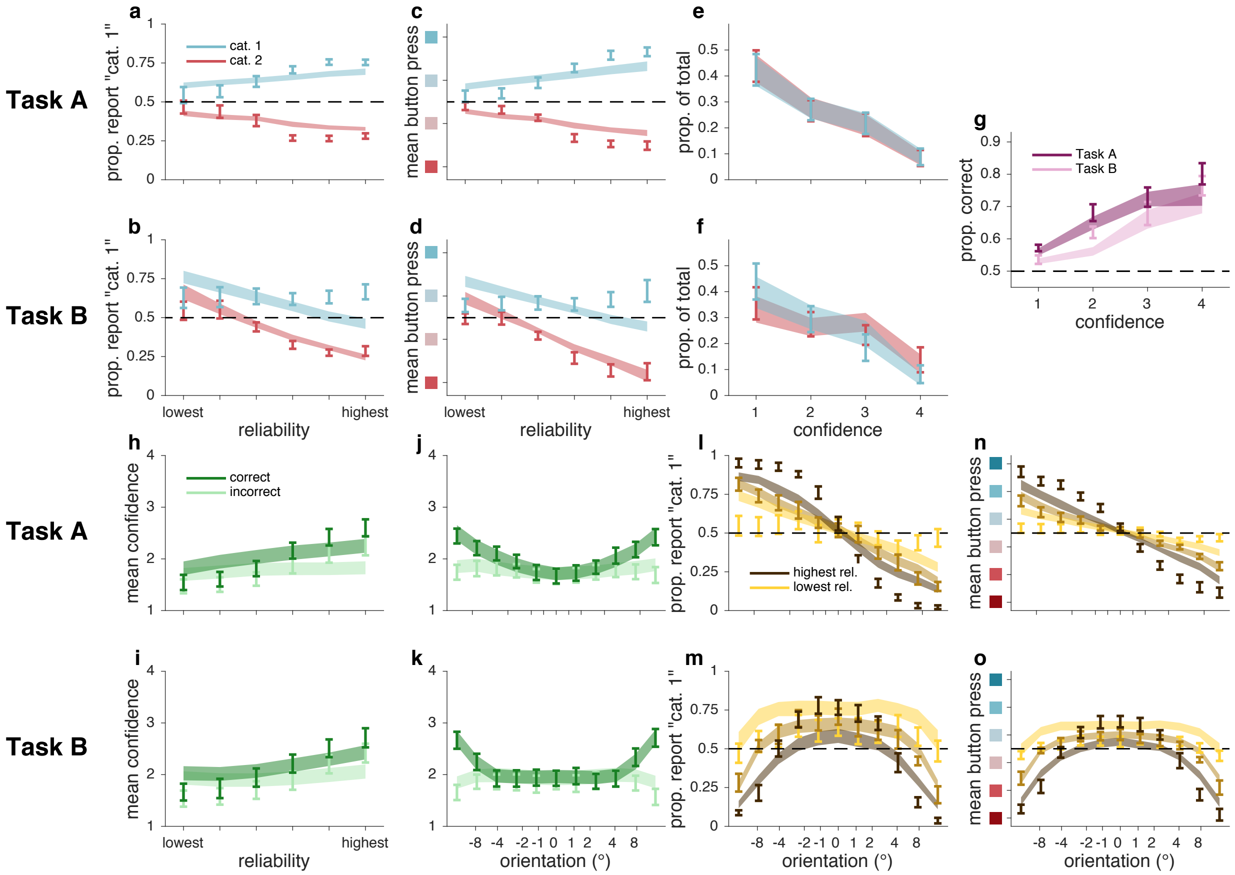

Figure S10: Orientation Estimation fits, as in **Figure 3**.

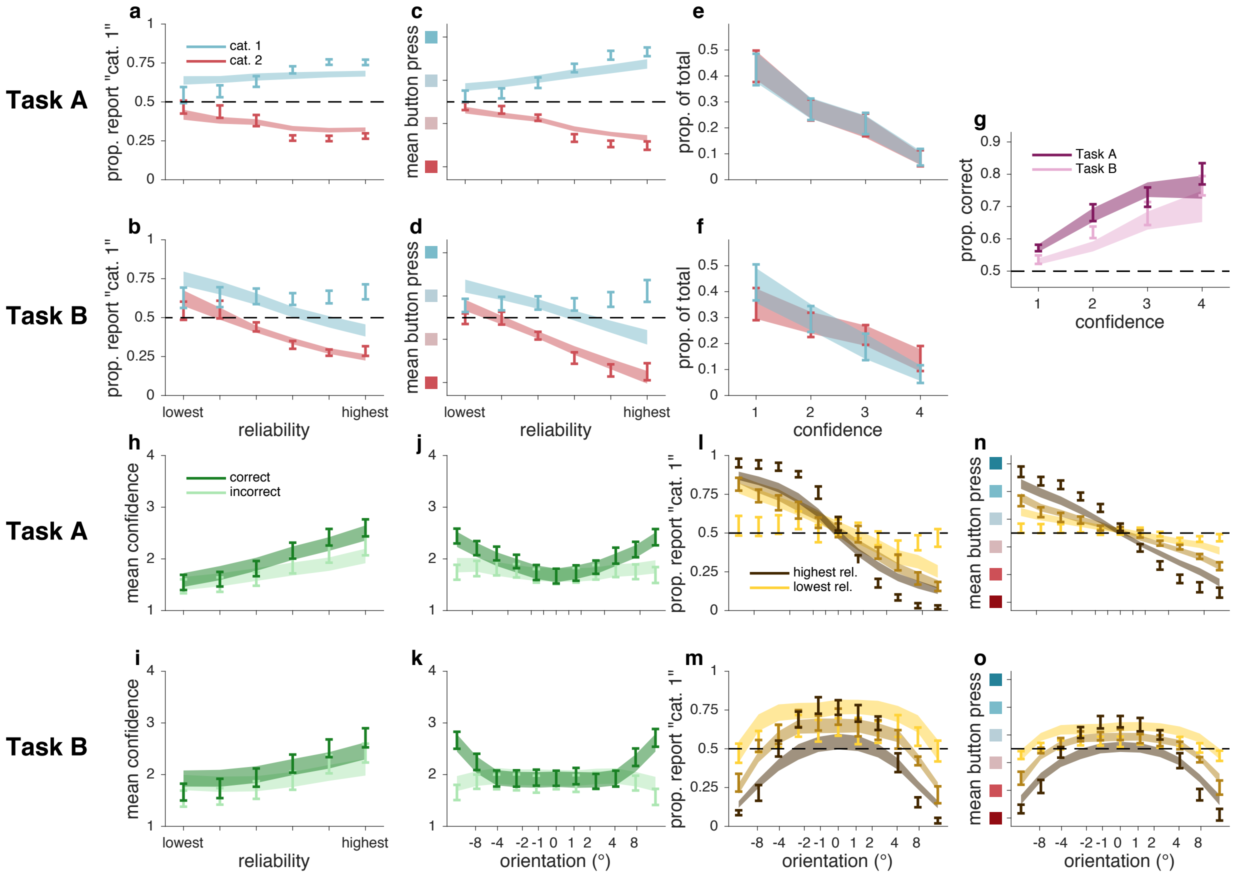

Figure S11: Linear Neural fits, as in **Figure 3**.

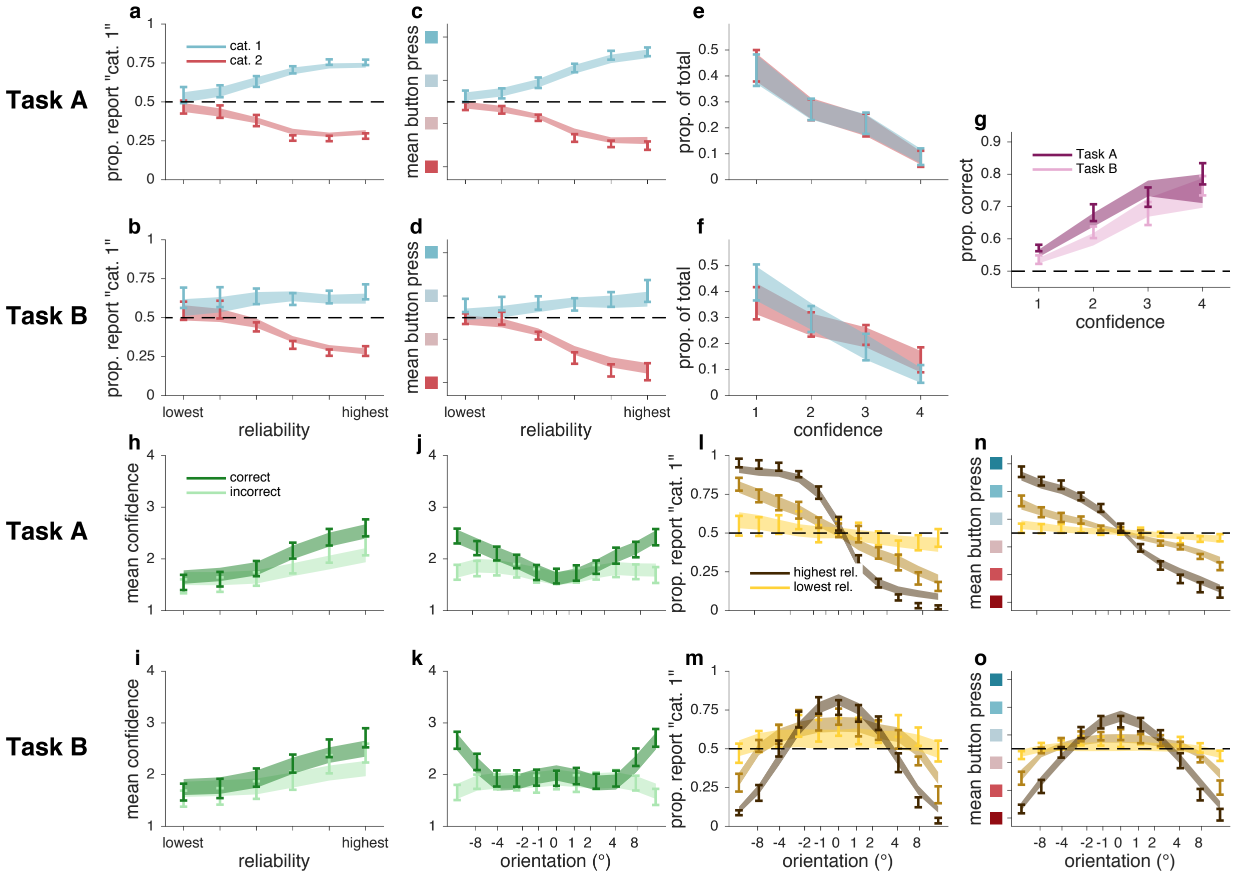

Figure S12: Lin fits, as in **Figure 3**.

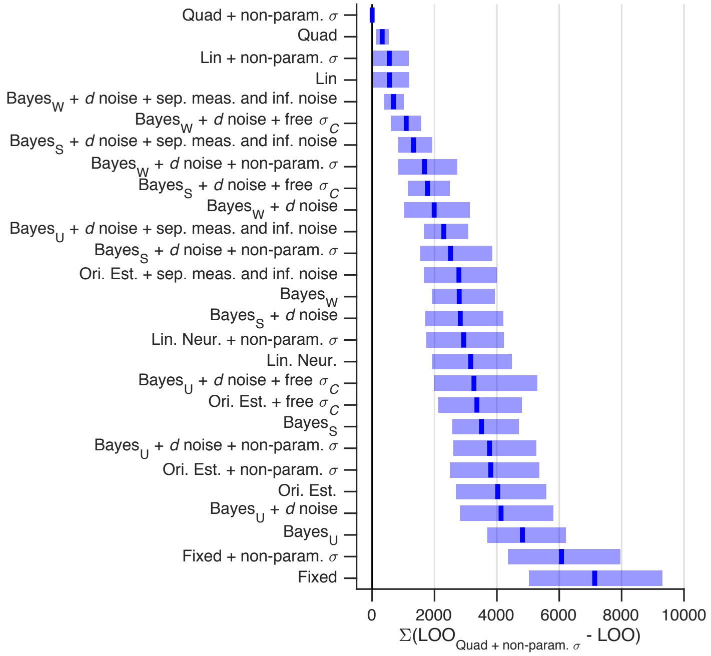

Figure S13: Model comparison, experiment 1. Models were fit jointly to Task A and B category and confidence responses. Medians and 95% CI of bootstrapped sums of LOO differences, relative to the best model. Higher values indicate worse fits.

		19 pars. Fixed	12 pars. Bayes$_U$	16 pars. Bayes$_S$	19 pars. Bayes$_W$	13 pars. Bayes$_U$ + d noise	17 pars. Bayes$_S$ + d noise	20 pars. Bayes$_W$ + d noise	19 pars. Ori. Est.	20 pars. Lin. Neur.	30 pars. Lin	30 pars. Quad	16 pars. Bayes$_U$ + d noise + non-param. σ	20 pars. Bayes$_S$ + d noise + non-param. σ	23 pars. Bayes$_W$ + d noise + non-param. σ	22 pars. Ori. Est. + non-param. σ	23 pars. Lin. Neur. + non-param. σ	33 pars. Lin + non-param. σ	33 pars. Quad + non-param. σ	22 pars. Fixed + non-param. σ	15 pars. Bayes$_U$ + d noise + free σ_C	19 pars. Bayes$_S$ + d noise + free σ_C	22 pars. Bayes$_W$ + d noise + free σ_C	21 pars. Ori. Est. + free σ_C	17 pars. Bayes$_U$ + d noise + sep. meas. and inf. noise	21 pars. Bayes$_S$ + d noise + sep. meas. and inf. noise	24 pars. Bayes$_W$ + d noise + sep. meas. and inf. noise
23 pars.	Ori. Est. + sep. meas. and inf. noise	−4356 [−5945, −2727]	−2042 [−3756, −457]	−701 [−2310, 596]	−3 [−1623, 1474]	−1347 [−3191, 222]	−3 [−1698, 1290]	802 [−784, 2223]	−1271 [−1836, −613]	−379 [−888, 186]	2226 [1419, 3120]	2459 [1384, 3640]	−946 [−2853, 571]	305 [−1532, 1634]	1109 [−468, 2587]	−1042 [−1639, −447]	−153 [−701, 413]	2218 [1411, 3136]	2790 [1653, 3984]	−3298 [−4486, −2152]	−525 [−2391, 1009]	1007 [42, 1957]	1708 [589, 2876]	−607 [−1133, 76]	472 [−726, 1685]	1464 [296, 2531]	2094 [1015, 3281]
24 pars.	Bayes$_W$ + d noise + sep. meas. and inf. noise	−6460 [−8426, −4486]	−4135 [−5389, −3054]	−2837 [−3750, −2030]	−2110 [−2954, −1451]	−3442 [−4979, −2233]	−2146 [−3244, −1241]	−1310 [−2175, −580]	−3351 [−4857, −2005]	−2486 [−3789, −1230]	126 [−492, 750]	362 [162, 602]	−3077 [−4407, −2038]	−1834 [−2890, −1085]	−995 [−1764, −388]	−3132 [−4639, −1802]	−2260 [−3579, −1009]	129 [−480, 743]	689 [389, 1017]	−5404 [−7167, −3772]	−2575 [−4535, −1366]	−1072 [−1730, −641]	−393 [−711, −115]	−2672 [−4062, −1494]	−1602 [−2390, −1007]	−636 [−1043, −347]	
21 pars.	Bayes$_S$ + d noise + sep. meas. and inf. noise	−5802 [−7674, −3955]	−3499 [−4427, −2629]	−2189 [−2894, −1565]	−1476 [−2216, −755]	−2814 [−3970, −1846]	−1501 [−2335, −814]	−667 [−1425, 36]	−2724 [−4031, −1403]	−1844 [−3000, −522]	760 [172, 1530]	1005 [579, 1549]	−2438 [−3427, −1641]	−1191 [−2016, −624]	−370 [−1019, 325]	−2519 [−3800, −1171]	−1645 [−2775, −325]	752 [168, 1526]	1331 [826, 1925]	−4754 [−6422, −3195]	−1948 [−3783, −905]	−439 [−773, −172]	248 [−140, 657]	−2041 [−3242, −871]	−971 [−1459, −544]		
17 pars.	Bayes$_U$ + d noise + sep. meas. and inf. noise	−4806 [−6759, −2996]	−2507 [−3204, −1980]	−1211 [−1932, −541]	−514 [−1354, 466]	−1840 [−2828, −1005]	−538 [−1444, 329]	298 [−645, 1325]	−1732 [−3025, −441]	−869 [−2064, 423]	1755 [1022, 2541]	1976 [1336, 2764]	−1453 [−2343, −791]	−236 [−1144, 609]	590 [−246, 1651]	−1543 [−2769, −234]	−661 [−1797, 642]	1741 [1029, 2544]	2303 [1651, 3088]	−3751 [−5537, −2211]	−1003 [−2673, −29]	525 [161, 945]	1210 [585, 1979]	−1062 [−2289, 112]			
21 pars.	Ori. Est. + free σ_C	−3756 [−5025, −2562]	−1456 [−3049, 22]	−110 [−1651, 1199]	571 [−970, 2109]	−757 [−2413, 579]	588 [−1004, 1747]	1370 [−100, 2776]	−671 [−940, −403]	212 [−337, 805]	2823 [1994, 3670]	3043 [1877, 4395]	−358 [−2089, 995]	891 [−831, 2182]	1669 [193, 3223]	−450 [−822, −85]	431 [−205, 1100]	2811 [1978, 3684]	3363 [2110, 4752]	−2699 [−3673, −1845]	45 [−1494, 1448]	1591 [639, 2624]	2275 [1176, 3602]				
22 pars.	Bayes$_W$ + d noise + free σ_C	−6059 [−7909, −4192]	−3751 [−4929, −2706]	−2434 [−3274, −1724]	−1710 [−2481, −1120]	−3043 [−4489, −1974]	−1742 [−2745, −984]	−908 [−1661, −318]	−2956 [−4401, −1654]	−2084 [−3410, −761]	518 [−39, 1156]	768 [399, 1144]	−2677 [−3916, −1764]	−1436 [−2402, −841]	−589 [−1260, −125]	−2740 [−4192, −1444]	−1868 [−3190, −546]	515 [−47, 1160]	1096 [607, 1571]	−5009 [−6718, −3406]	−2199 [−3914, −1144]	−678 [−1261, −291]					
19 pars.	Bayes$_S$ + d noise + free σ_C	−5342 [−7118, −3637]	−3056 [−3956, −2240]	−1726 [−2524, −1084]	−1044 [−1888, −126]	−2361 [−3442, −1517]	−1041 [−1953, −377]	−234 [−1088, 650]	−2268 [−3377, −1157]	−1406 [−2436, −253]	1220 [656, 1856]	1451 [894, 2145]	−1969 [−2982, −1262]	−736 [−1688, −104]	61 [−690, 985]	−2072 [−3135, −932]	−1191 [−2191, −32]	1206 [662, 1861]	1779 [1144, 2487]	−4291 [−5905, −2865]	−1535 [−3085, −678]						
15 pars.	Bayes$_U$ + d noise + free σ_C	−3737 [−5664, −2089]	−1549 [−2740, 46]	−202 [−1392, 1367]	495 [−765, 2112]	−847 [−2103, 660]	449 [−798, 2020]	1296 [34, 2933]	−720 [−2280, 1032]	136 [−1487, 2235]	2736 [1604, 4310]	2960 [1706, 4864]	−502 [−1809, 1147]	752 [−598, 2436]	1600 [337, 3343]	−536 [−2021, 1277]	345 [−1261, 2460]	2739 [1604, 4248]	3289 [1961, 5318]	−2758 [−4820, −565]							
22 pars.	Fixed + non-param. σ	−1015 [−2165, −209]	1211 [−620, 3272]	2551 [826, 4356]	3269 [1479, 5190]	1910 [72, 3813]	3270 [1468, 4897]	4073 [2379, 5781]	2025 [1280, 2914]	2921 [2044, 3904]	5539 [4089, 7019]	5767 [4106, 7522]	2306 [386, 4270]	3574 [1698, 5324]	4378 [2598, 6266]	2259 [1384, 3301]	3141 [2140, 4228]	5532 [4088, 7029]	6097 [4323, 7901]								
33 pars.	Quad + non-param. σ	−7161 [−9294, −4986]	−4818 [−6203, −3697]	−3521 [−4673, −2545]	−2795 [−3913, −1923]	−4136 [−5848, −2777]	−2837 [−4189, −1685]	−1996 [−3163, −1026]	−4049 [−5567, −2649]	−3180 [−4493, −1931]	−555 [−1198, −19]	−325 [−535, −141]	−3760 [−5309, −2594]	−2525 [−3864, −1557]	−1676 [−2730, −839]	−3830 [−5348, −2504]	−2950 [−4252, −1743]	−550 [−1181, −15]									
33 pars.	Lin + non-param. σ	−6602 [−8290, −4809]	−4249 [−5664, −3167]	−2920 [−4204, −2108]	−2217 [−3462, −1323]	−3549 [−5189, −2402]	−2238 [−3637, −1362]	−1401 [−2652, −573]	−3490 [−4532, −2508]	−2612 [−3524, −1691]	−5 [−16, 8]	229 [−234, 773]	−3171 [−4764, −2142]	−1945 [−3376, −1111]	−1102 [−2291, −243]	−3269 [−4337, −2329]	−2396 [−3326, −1474]										
23 pars.	Lin. Neur. + non-param. σ	−4180 [−5901, −2591]	−1863 [−3689, −357]	−529 [−2321, 828]	151 [−1673, 1764]	−1172 [−3186, 358]	155 [−1768, 1536]	966 [−876, 2526]	−1101 [−1639, −572]	−222 [−381, −76]	2388 [1475, 3310]	2612 [1437, 3898]	−768 [−2847, 708]	469 [−1586, 1905]	1268 [−548, 2929]	−886 [−1362, −434]											
22 pars.	Ori. Est. + non-param. σ	−3295 [−4793, −1910]	−993 [−2755, 523]	351 [−1428, 1801]	1018 [−782, 2782]	−290 [−2170, 1090]	1042 [−803, 2396]	1823 [84, 3478]	−217 [−405, −47]	662 [272, 1105]	3259 [2324, 4317]	3487 [2225, 4992]	113 [−1858, 1515]	1346 [−645, 2818]	2119 [382, 3940]												
23 pars.	Bayes$_W$ + d noise + non-param. σ	−5436 [−7248, −3675]	−3113 [−4280, −2194]	−1824 [−2424, −1293]	−1114 [−1377, −858]	−2401 [−3750, −1551]	−1120 [−1794, −646]	−306 [−505, −120]	−2355 [−4056, −683]	−1487 [−3074, 259]	1104 [264, 2301]	1351 [686, 2291]	−2052 [−3165, −1304]	−839 [−1353, −449]													
20 pars.	Bayes$_S$ + d noise + non-param. σ	−4586 [−6336, −2866]	−2276 [−3132, −1568]	−971 [−1397, −613]	−275 [−715, 262]	−1581 [−2477, −986]	−293 [−522, −96]	535 [131, 1006]	−1548 [−3013, 304]	−680 [−2096, 1279]	1945 [1114, 3372]	2186 [1351, 3450]	−1221 [−1854, −748]														
16 pars.	Bayes$_U$ + d noise + non-param. σ	−3313 [−5206, −1581]	−1030 [−1503, −666]	251 [−264, 910]	940 [243, 2018]	−361 [−675, −135]	941 [441, 1478]	1763 [981, 2774]	−302 [−1745, 1556]	570 [−962, 2571]	3188 [2125, 4785]	3449 [2328, 4882]															
30 pars.	Quad	−6832 [−8810, −4810]	−4494 [−5853, −3399]	−3189 [−4257, −2329]	−2463 [−3458, −1757]	−3799 [−5455, −2521]	−2509 [−3784, −1487]	−1667 [−2698, −858]	−3714 [−5206, −2408]	−2840 [−4117, −1671]	−233 [−790, 244]																
30 pars.	Lin	−6596 [−8279, −4812]	−4246 [−5658, −3164]	−2916 [−4196, −2106]	−2213 [−3458, −1318]	−3545 [−5177, −2401]	−2234 [−3628, −1364]	−1398 [−2644, −571]	−3485 [−4519, −2515]	−2607 [−3513, −1694]																	
20 pars.	Lin. Neur.	−3959 [−5540, −2506]	−1645 [−3429, −125]	−309 [−2049, 1036]	373 [−1405, 1978]	−953 [−2922, 565]	381 [−1490, 1720]	1188 [−588, 2704]	−878 [−1325, −450]																		
19 pars.	Ori. Est.	−3080 [−4409, −1862]	−779 [−2498, 767]	564 [−1154, 1997]	1239 [−480, 2938]	−77 [−1881, 1321]	1261 [−526, 2571]	2041 [385, 3623]																			
20 pars.	Bayes$_W$ + d noise	−5134 [−6764, −3509]	−2824 [−3938, −1866]	−1517 [−2086, −999]	−804 [−1134, −510]	−2105 [−3353, −1281]	−819 [−1369, −441]																				
17 pars.	Bayes$_S$ + d noise	−4294 [−5857, −2733]	−1983 [−2814, −1263]	−662 [−1150, −307]	14 [−497, 691]	−1295 [−2051, −734]																					
13 pars.	Bayes$_U$ + d noise	−2944 [−4744, −1329]	−657 [−1217, −214]	607 [−8, 1501]	1301 [514, 2625]																						
19 pars.	Bayes$_W$	−4312 [−6151, −2610]	−1987 [−3104, −1204]	−695 [−1222, −353]																							
16 pars.	Bayes$_S$	−3595 [−5323, −1995]	−1298 [−1951, −747]																								
12 pars.	Bayes$_U$	−2265 [−4253, −498]																									

Table S1: Cross comparison of all models in **Figure S13**. Cells indicate medians and 95% CI of bootstrapped summed LOO score differences. A negative median indicates that the model in the corresponding row had a higher score (better fit) than the model in the corresponding column. For readability, see *table_S1.xls*. See *model_parameters.xls* for the parameters of each model.

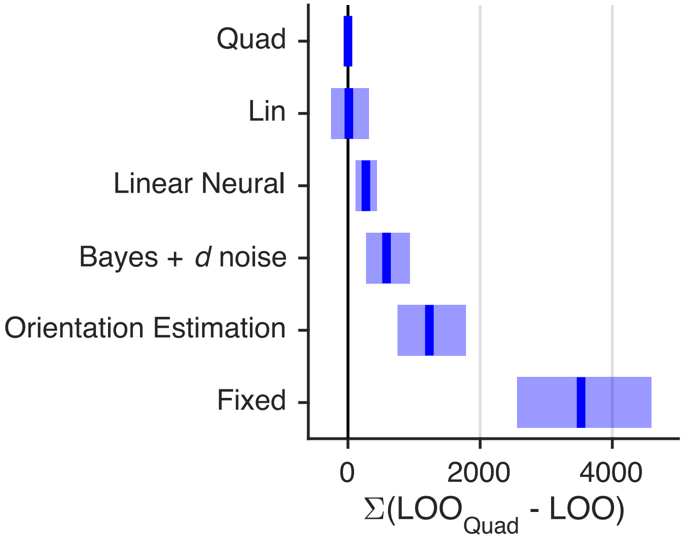

Figure S14: Model comparison, experiment 1. Models were fit to Task A category and confidence responses. Medians and 95% CI of bootstrapped sums of LOO differences, relative to the best model. Higher values indicate worse fits.

		12 pars.	13 pars.	12 pars.	13 pars.	16 pars.
		Fixed	Bayes + d noise	Ori. Est.	Lin. Neur.	Lin
16 pars.	Quad	-3534 $[-4552, -2529]$	-581 $[-938, -278]$	-1241 $[-1798, -767]$	-270 $[-436, -117]$	-14 $[-325, 246]$
16 pars.	Lin	-3532 $[-4353, -2651]$	-572 $[-799, -339]$	-1232 $[-1566, -863]$	-259 $[-609, 124]$	
13 pars.	Lin. Neur.	-3255 $[-4343, -2231]$	-313 $[-724, 75]$	-972 $[-1599, -412]$		
12 pars.	Ori. Est.	-2302 $[-2881, -1705]$	651 $[425, 885]$			
13 pars.	Bayes + d noise	-2956 $[-3723, -2163]$				

Table S2: Cross comparison of all models in **Figure S14**. See **Table S1** caption.

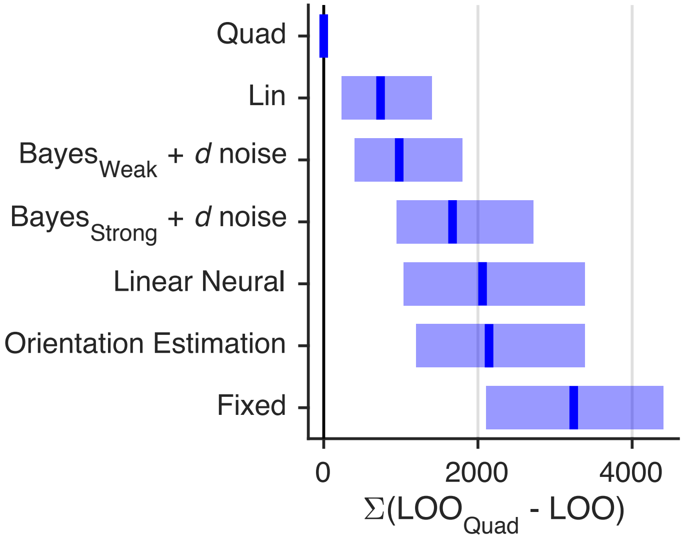

Figure S15: Model comparison, experiment 1. Models were fit to Task B category and confidence responses. Medians and 95% CI of bootstrapped sums of LOO differences, relative to the best model. Higher values indicate worse fits.

		15 pars.	13 pars.	16 pars.	15 pars.	16 pars.	22 pars.
		Fixed	Bayes_S + d noise	Bayes_W + d noise	Ori. Est.	Lin. Neur.	Lin
22 pars.	Quad	-3234 $[-4390, -2099]$	-1664 $[-2698, -958]$	-978 $[-1756, -406]$	-2156 $[-3352, -1192]$	-2060 $[-3368, -1037]$	-744 $[-1387, -224]$
22 pars.	Lin	-2480 $[-3323, -1645]$	-919 $[-1788, -279]$	-232 $[-900, 346]$	-1415 $[-2439, -439]$	-1326 $[-2442, -337]$	
16 pars.	Lin. Neur.	-1117 $[-2093, -349]$	421 $[-1095, 1689]$	1106 $[-374, 2583]$	-80 $[-222, 62]$		
15 pars.	Ori. Est.	-1043 $[-1962, -273]$	502 $[-934, 1693]$	1184 $[-202, 2588]$			
16 pars.	Bayes_W + d noise	-2230 $[-3239, -1307]$	-691 $[-1082, -390]$				
13 pars.	Bayes_S + d noise	-1534 $[-2425, -634]$					

Table S3: Cross comparison of all models in **Figure S15**. See **Table S1** caption.

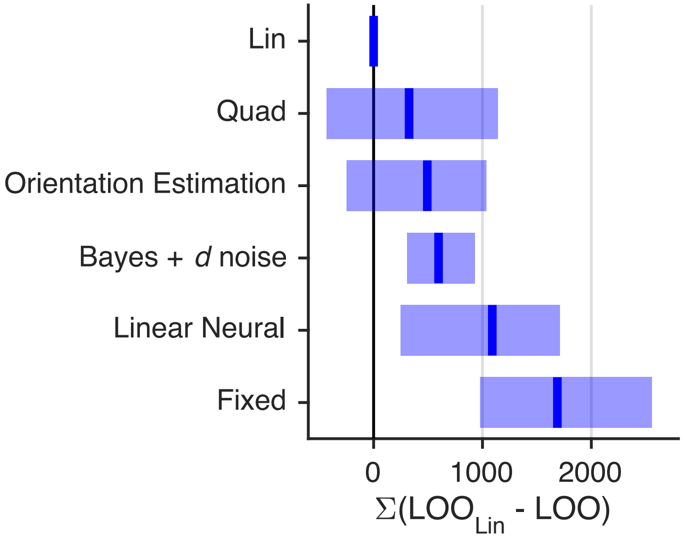

Figure S16: Model comparison, experiment 1. Models were fit jointly to Task A and B category choices. Medians and 95% CI of bootstrapped sums of LOO differences, relative to the best model. Higher values indicate worse fits.

		8 pars.	9 pars.	8 pars.	9 pars.	10 pars.
		Fixed	Bayes + d noise	Ori. Est.	Lin. Neur.	Lin
10 pars.	Quad	-1319 $[-2541, -611]$	-236 $[-1072, 358]$	-154 $[-772, 613]$	-729 $[-1365, -38]$	323 $[-423, 1127]$
10 pars.	Lin	-1690 $[-2534, -976]$	-595 $[-927, -311]$	-492 $[-1023, 238]$	-1087 $[-1690, -245]$	
9 pars.	Lin. Neur.	-591 $[-2068, 460]$	486 $[-504, 1211]$	591 $[406, 789]$		
8 pars.	Ori. Est.	-1190 $[-2614, -144]$	-114 $[-1026, 579]$			
9 pars.	Bayes + d noise	-1095 $[-1657, -629]$				

Table S4: Cross comparison of all models in **Figure S16**. See **Table S1** caption.

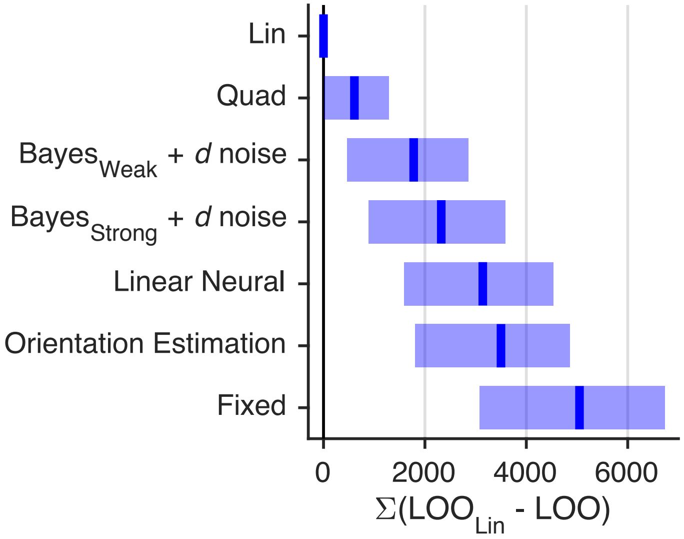

Figure S17: Model comparison, experiment 1. Noise parameters were fit to Task A category choices and then fixed during the fitting of Task B category and confidence responses. Medians and 95% CI of bootstrapped sums of LOO differences, relative to the best model. Higher values indicate worse fits.

		15 pars.	13 pars.	16 pars.	15 pars.	16 pars.	22 pars.
		Fixed	Bayes_S + d noise	Bayes_W + d noise	Ori. Est.	Lin. Neur.	Lin
22 pars.	Quad	$-4367\ [-6304, -2391]$	$-1670\ [-3268, -39]$	$-1135\ [-2501, 333]$	$-2836\ [-4544, -1122]$	$-2449\ [-4200, -969]$	$606\ [6, 1269]$
22 pars.	Lin	$-5016\ [-6727, -3090]$	$-2303\ [-3578, -921]$	$-1773\ [-2845, -451]$	$-3497\ [-4860, -1817]$	$-3127\ [-4549, -1575]$	
16 pars.	Lin. Neur.	$-1837\ [-3566, -378]$	$846\ [-609, 2092]$	$1386\ [-13, 2732]$	$-345\ [-933, 262]$		
15 pars.	Ori. Est.	$-1498\ [-2877, -420]$	$1184\ [23, 2129]$	$1724\ [575, 2808]$			
16 pars.	Bayes_W + d noise	$-3257\ [-3965, -2494]$	$-533\ [-920, -283]$				
13 pars.	Bayes_S + d noise	$-2704\ [-3351, -2027]$					

Table S5: Cross comparison of all models in **Figure S17**. See **Table S1** caption.

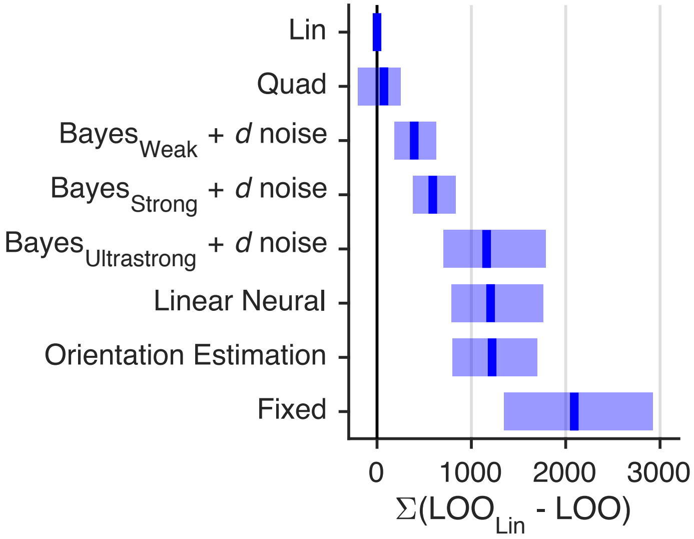

Figure S18: Model comparison, experiment 2. Models were fit jointly to Task A and B category and confidence responses. Medians and 95% CI of bootstrapped sums of LOO differences, relative to the best model. Higher values indicate worse fits.

		19 pars.	13 pars.	17 pars.	20 pars.	19 pars.	20 pars.	30 pars.
		Fixed	Bayes$_\text{U}$ + d noise	Bayes$_\text{S}$ + d noise	Bayes$_\text{W}$ + d noise	Ori. Est.	Lin. Neur.	Lin
30 pars.	Quad	-2014 $[-3036, -1186]$	-1096 $[-1807, -530]$	-523 $[-893, -220]$	-331 $[-562, -109]$	-1136 $[-1815, -638]$	-1124 $[-1922, -613]$	74 $[-195, 252]$
30 pars.	Lin	-2095 $[-2889, -1344]$	-1160 $[-1780, -694]$	-589 $[-841, -375]$	-396 $[-622, -186]$	-1218 $[-1680, -791]$	-1205 $[-1757, -785]$	
20 pars.	Lin. Neur.	-876 $[-1401, -395]$	55 $[-711, 693]$	623 $[99, 1184]$	801 $[253, 1491]$	-7 $[-219, 216]$		
19 pars.	Ori. Est.	-872 $[-1297, -487]$	56 $[-561, 574]$	620 $[218, 1089]$	813 $[356, 1394]$			
20 pars.	Bayes$_\text{W}$ + d noise	-1678 $[-2542, -1007]$	-767 $[-1262, -386]$	-190 $[-363, -82]$				
17 pars.	Bayes$_\text{S}$ + d noise	-1490 $[-2210, -886]$	-565 $[-1032, -266]$					
13 pars.	Bayes$_\text{U}$ + d noise	-907 $[-1572, -365]$						

Table S6: Cross comparison of all models in **Figure S18**. See **Table S1** caption.

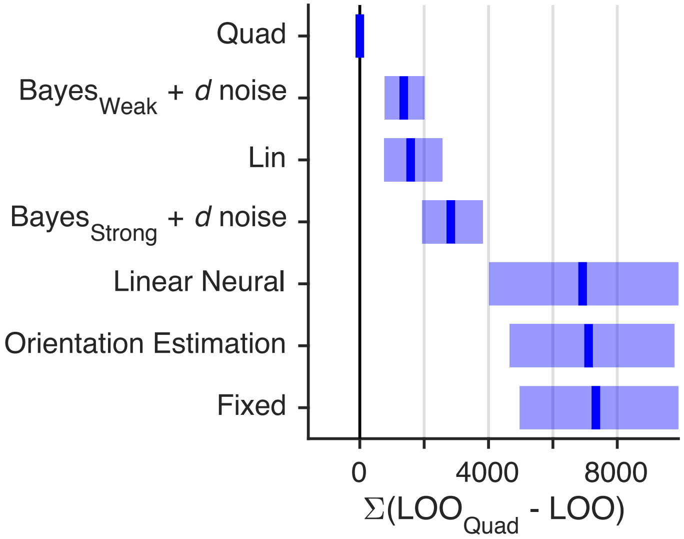

Figure S19: Model comparison, experiment 3. Models were fit to Task B category and confidence responses. Medians and 95% CI of bootstrapped sums of LOO differences, relative to the best model. Higher values indicate worse fits.

		15 pars.	13 pars.	16 pars.	15 pars.	16 pars.	22 pars.
		Fixed	Bayes_S + d noise	Bayes_W + d noise	Ori. Est.	Lin. Neur.	Lin
22 pars.	Quad	-7326 $[-9955, -4905]$	-2833 $[-3807, -1926]$	-1361 $[-2022, -777]$	-7120 $[-9838, -4636]$	-6902 $[-10376, -3981]$	-1577 $[-2562, -750]$
22 pars.	Lin	-5759 $[-7866, -3694]$	-1240 $[-2567, 65]$	226 $[-812, 1246]$	-5530 $[-7707, -3539]$	-5337 $[-8191, -2846]$	
16 pars.	Lin. Neur.	-450 $[-1535, 1290]$	4114 $[733, 7796]$	5552 $[2338, 9135]$	-214 $[-1176, 1256]$		
15 pars.	Ori. Est.	-256 $[-841, 423]$	4311 $[1432, 7134]$	5727 $[3067, 8527]$			
16 pars.	Bayes_W + d noise	-5967 $[-8702, -3369]$	-1454 $[-2179, -835]$				
13 pars.	Bayes_S + d noise	-4505 $[-7282, -1816]$					

Table S7: Cross comparison of all models in **Figure S19**. See **Table S1** caption.

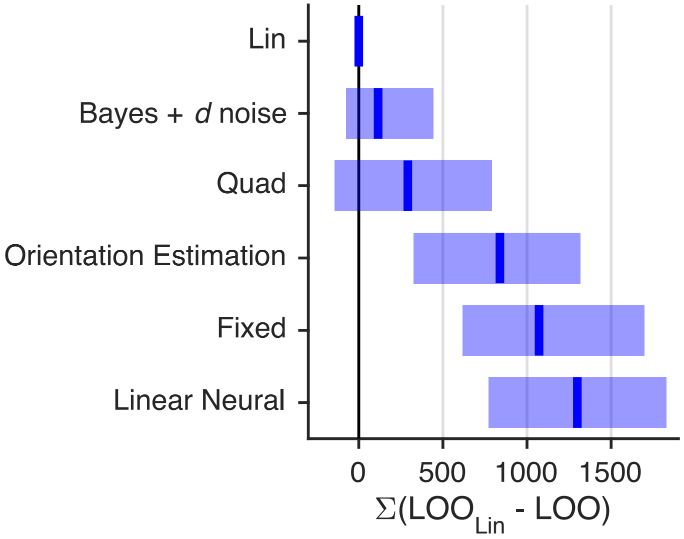

Figure S20: Model comparison, experiment 3. Models were fit to Task B category choices. Medians and 95% CI of bootstrapped sums of LOO differences, relative to the best model. Higher values indicate worse fits.

		7 pars.	8 pars.	7 pars.	8 pars.	8 pars.
		Fixed	Bayes + d noise	Ori. Est.	Lin. Neur.	Lin
8 pars.	Quad	-777 $[-1361, -359]$	162 $[-290, 670]$	-531 $[-1059, -23]$	-988 $[-1526, -549]$	290 $[-138, 793]$
8 pars.	Lin	-1084 $[-1675, -619]$	-117 $[-436, 76]$	-830 $[-1317, -334]$	-1294 $[-1825, -778]$	
8 pars.	Lin. Neur.	215 $[-566, 827]$	1174 $[535, 1772]$	457 $[254, 685]$		
7 pars.	Ori. Est.	-255 $[-987, 369]$	707 $[119, 1259]$			
8 pars.	Bayes + d noise	-964 $[-1290, -663]$				

Table S8: Cross comparison of all models in **Figure S20**. See **Table S1** caption.