Computational mechanisms underlying human confidence reports

by

William T. Adler

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Center for Neural Science

New York University

January 2018

Wei Ji Ma

I think it's much more interesting to live not knowing than to have answers which might be wrong. I have approximate answers, and possible beliefs, and different degrees of certainty about different things, but I'm not absolutely sure of anything, and there are many things I don't know anything about, such as whether it means anything to ask why we're here, and what the question might mean... But I don't have to know an answer. I don't feel frightened by not knowing things, by being lost in a mysterious universe without having any purpose, which is the way it really is, as far as I can tell, possibly.

Richard Feynman

Dedication

To Maw, who asked questions

Acknowledgements

First and foremost, I want to thank Wei Ji Ma. Weiji is the best graduate school mentor that I could have asked for. I feel lucky to have joined Weiji's lab, in part because it was so unplanned (he was not yet NYU faculty when I applied here), and in part because he allowed me to join even though I knew next to nothing about probability. Weiji has always pushed me to be extremely rigorous in forming and answering questions, and has helped me become a more precise thinker and writer. He has been a friend and an ally in political matters inside and outside of NYU. I wish him all the best in his new journey of fatherhood, and I know that he and Ting will make great parents.

Rachel Denison has been an excellent collaborator on the project presented in Chapter 4. Collaborating with her has shown me how much more fun and productive it is to work with someone I admire than to work alone. Thanks to Marisa Carrasco for knowing everything there is to know about attention. Roshni Lulla and Gordon Bill were a great help on this project, collecting data and having thoughtful discussions.

I want to thank my committee, Eero Simoncelli, Roozbeh Kiani, Marisa Carrasco, and my outside reader Angela Yu, for helping to shape this work.

Steve Fleming and Joaquín Navajas provided helpful feedback on the work presented in Chapter 2.

Thank you to the National Science Foundation and all political actors that enable taxpayer-funded research, without which most science would not be possible. This dissertation was based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1342536.

I want to highlight a few of my brilliant colleagues in the Ma lab. Luigi Acerbi has been one of the most important figures of my graduate education. I learned something new

every time I talked to Luigi, and his facility and enthusiasm for modeling were contagious. Aspen Yoo has been one of my best friends, and a great support, ever since she mysteriously appeared in the lab one day without any announcement or introduction by Weiji. I have been indebted to Bas van Opheusden ever since he stayed up with me until 3 a.m. our first year of graduate school, helping me write my first log likelihood function. Andra Mihali has always provided helpful comments on manuscripts and talks, as well as ibuprofen for headaches. Emin Orhan has provided various mathematical assistance over the years, and the analysis in Chapter 5 was very much inspired by his work.

I want to thank everyone who has been with the Scientist Action and Advocacy Network (ScAAN) since the beginning. I have been so impressed by your enthusiasm for combining science with policy, and by the hard work you have put into our projects. You all have helped me to expand my conception of what a scientific career can be. Many people have provided meaningful contributions to ScAAN projects, but I especially want to thank Weiji, Jenn, Milenna, Kristina, David, Silvia, Ali, Billy, Stephen, Andra, and EG, for their consistent participation, enthusiasm, and leadership. Thanks to the department, especially Amala, for supporting us.

Thank you to my New York friends in and outside of the department, especially Silvia Lopez Guzman, Corey Ziemba, and Alex Gibson.

I want to thank my supportive family, especially my parents Martha and David, who taught me to love learning. Thank you to my sisters Zoe and Rachel; my soon-to-be parents-in-law Lisa and Philip; my Hags family Tania, Ariel, Charlie, and Robbie; and Jolene and DeWitt Clinton, who only occasionally spill coffee on my keyboard.

Finally, I want to thank Sarah. Over the last eight years, we have watched each other grow more fully into ourselves and into our relationship. I treasure the time that we have spent together, and I am thankful that we will have so much more of it.

Preface

Each of Chapters 2 to 4 represents work from a manuscript that has been previously posted as a preprint on *bioRxiv*, with myself as first author and Wei Ji Ma as senior author.

On the manuscript associated with Chapter 4, Rachel N. Denison and I are co-first authors, and Marisa Carrasco is also an author.

Chapter 5 was included in an earlier version of the manuscript associated with Chapter 3, and is in that article's version history on *bioRxiv*. It has since been removed from the manuscript, and is unlikely to see publication outside of this dissertation.

All human behavior data, as well as all code used for the work presented here, is available at github.com/wtadler/confidence.

Abstract

A perceptual decision is often accompanied by a subjective feeling of confidence. Because humans are able to easily report this feeling in a laboratory setting, confidence reports have long been objects of study. However, the computations underlying confidence reports are not well understood.

It has been proposed that confidence in categorization tasks should be defined as the observer's estimated probability of being correct. This definition extends Bayesian decision theory so that it describes confidence reports as well as decisions. Although this definition is elegant, the notion that confidence reports are Bayesian is a hypothesis rather than an established fact. In this dissertation, our aim is to test that hypothesis, which we call the Bayesian confidence hypothesis (BCH).

We find that a proposed approach to determining the computational origins of confidence is flawed. Some authors have proposed that one way to determine whether confidence is Bayesian is to derive qualitative signatures of Bayesian confidence, and then see whether they are present in behavioral or neural data. We analyze some of these proposed signatures and find that they are less useful than they might have seemed. Specifically, they are neither necessary nor sufficient signatures of Bayesian confidence, which means that observation of (or failure to observe) these signatures provides an uncertain amount of evidence for (or against) the BCH. There has been a confusion in the literature about a second possible signature. We find no evidence that this second signature is ever expected under Bayesian confidence. Finally, the application of these signatures is a qualitative exercise because it may not always be clear whether data displays a signature, especially noisy data. Our analysis of the signatures leads us to conclude that the most powerful way to test the BCH is by using quantitative model comparison.

We test human subjects on a set of binary categorization tasks designed to distinguish Bayesian models of confidence from other plausible models. In all experiments, the primary variable of interest to the observer was the orientation of a stimulus.

In one set of experiments, we induce sensory uncertainty by manipulating properties of the stimulus, such as contrast. We find that subjects take their sensory uncertainty into account, and that confidence appears qualitatively Bayesian. Quantitatively, however, heuristic models provide a much better fit to the data. Our conclusions are robust to variants of both the experiment and the Bayesian models.

In another experiment, we induce sensory uncertainty by manipulating the subjects' attention. As in the previous set of experiments, we find that confidence reports are qualitatively Bayesian. In this experiment, we are unable to distinguish the Bayesian model from the heuristic models.

Finally, we describe an exploratory analysis intended to explain why confidence reports might not be Bayesian. We trained feedforward neural networks on our tasks as if they were naïve human subjects and fit our behavioral models to the data produced by these trained networks. We find that the same heuristic models that fit our human data well also fit the network-produced data. We suggest a future research program in which neural network behavior is compared to human behavior on the basis of model rankings.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We often have a sense of confidence in our percepts and in the decisions that we make based on those percepts. For instance, imagine that you see a person some distance away, and you have a feeling that the person is your friend. If the person is far away, you may have relatively low confidence in your belief that the person is your friend. But as you approach the person, you may feel an increasing sense of confidence, which eventually crosses some threshold, leading you to decide to wave “hello.” In this case, you derived from your retinal input an image that led to a belief about the person’s identity, as well as information about the uncertainty associated with that image. Through some unknown process, you combined the two pieces of information, waving only when confident that you would avoid the humiliation of waving at a stranger.

Such a “feeling of knowing” (Brown, 1991; Meyniel et al., 2015) may help humans and nonhuman animals make better decisions. When driving through a storm, if someone has low confidence about the speed and distance of the car in front of you, he may drive more conservatively. A radiologist’s confidence in her classification of a tumor as malignant or benign may partially determine her patient’s course of treatment. A sense of confidence may improve learning (Meyniel et al., 2015). Having an internal sense of confidence might also allow observers to update decision-making strategies in response to feedback: if a decision made with high confidence turns out to be incorrect, it might be time to change strategies (Purcell and Kiani, 2016).

The ability to express confidence plays an important role in group decision-making. A group often makes better decisions than even the best individual in a group (Frith and Frith, 2012). Some results suggest that groups achieve this “two heads are better than one” effect by optimally weighting each group member’s decision by their confidence (Bahrami et al., 2010; Koriat, 2012). In some situations, a group may also adopt a simple “confidence heuristic” (Thomas and McFadyen, 1995) strategy in which it selects the decision of its most confident member.[I]

In addition to being useful to organisms, confidence reports can also be used by experimenters as a tool for quantifying an organism’s ability to understand its own cognition (i.e., their metacognitive abilities). A number of metrics have been developed for measuring the ability to distinguish between one’s correct and incorrect judgments (Fleming and Lau, 2014; Maniscalco and Lau, 2012). These metrics can be used to compare the metacognitive abilities of different populations or species. For instance, older people have been shown to have reduced metacognitive abilities (Palmer et al., 2014). This might explain why confidence and performance become increasingly dissociated as we age, as evidenced by the fact that most older drivers rate themselves as good drivers regardless of their history of crashes (Ross et al., 2012).

1.1 Techniques for measuring confidence

There are many techniques that allow researchers to measure confidence in humans and nonhuman animals, which have been thoroughly reviewed in Kepecs and Mainen (2012).

[I] This strategy, however, can be harmful to group decision-making for difficult decisions or in cases where group members have different levels of task performance (Bang et al., 2014; Koriat, 2012).

These methods can include post-decision wagering (Persaud et al., 2007), in which a subject places a bet on their choice being correct. There are other methods that can be used to collect confidence ratings from nonhuman animals, such as offering a “sure bet” option, in which the animal opts not to make a choice between categories, but to instead receive a smaller but certain reward. This method has been used to collect confidence ratings from monkeys (Kiani and Shadlen, 2009; Smith et al., 1997), pigeons (Sutton and Shettleworth, 2008), rats (Foote and Crystal, 2007), and dolphins (Smith et al., 1995). Another option is to measure the amount of time that an animal is willing to wait for a reward before restarting a trial; an animal can maximize its reward rate by waiting longer on trials where it is more confident (Kepecs et al., 2008).

A common problem in experiments using confidence reports such as these is that it is not always clear that the chosen technique truly measures the observer’s subjective feeling of perceptual confidence.[II] For instance, in experiments that use post-decision wagering, observers may simply learn to maximize their reward. In this case, experimenters are training the observer to report a particular form of confidence. Similarly, the small reward associated with a “sure bet” choice may itself guide behavior, making it possible that ostensible “confidence” reports merely reflect some reward-conditioned behavior rather than any subjective feeling of confidence (Smith et al., 2008). There are some ways around these problems, which are particularly critical for researchers studying confidence in nonhumans (Hampton, 2009; Kepecs and Mainen, 2012; Smith et al., 2008). However, researchers studying human confidence need not concern themselves with these issues—to get a confidence report from a human, we just have to ask.

The most straightforward paradigm for eliciting confidence reports, which to this point has

[II] We sidestep the philosophical question of whether perceptual states (categorical or otherwise) are accompanied by a subjective feeling of confidence (Denison, 2017; Morrison, 2016, 2017), and assume that they are.

only been used in humans, is to ask subjects to explicitly rate their confidence on an integer scale. This technique has been in use for over a century; the first known psychophysical exploration of confidence ratings was in 1884, by Peirce and Jastrow (1884). The authors conducted a experiment that is roughly similar to binary perceptual categorization experiments conducted by researchers today and to those described in this dissertation. They had subjects hold two weights and say which one was heavier, and also state their confidence that their judgment was correct. They found, curiously, that for choices where the subject reported zero confidence, the subject still performed above chance.[III] Theirs is the first recorded experiment to have probed the interaction between physical stimulus, choice, and subjective confidence reports. Since then, this rating scale method has been used in dozens of experiments.

Recent work has focused on identifying brain regions and neural mechanisms responsible for the computation of confidence in humans (Fleming and Dolan, 2012; Fleming et al., 2010; Rutishauser et al., 2015), nonhuman primates (Fetsch et al., 2014; Kiani and Shadlen, 2009; Komura et al., 2013), and rodents (Kepecs et al., 2008). It has been argued that the search for neural correlates of confidence would be more fruitful if researchers were equipped with a strong model of confidence and knew what kind of signals to look for in neural activity (Kepecs and Mainen, 2012; Pouget et al., 2016).

1.2 Models of confidence

Despite the long history of collecting explicit confidence ratings, relatively little work has been done to understand the computations that transform a sensory measurement into a confidence rating. Confidence ratings have been frequently used as a tool for computing ROC (Receiver Operating Characteristic) curves, used to measure sensitivity, or for measuring

[III] Note, however, that the subjects were the experimenters themselves (as was the norm in nineteenth-century experimental psychology), and that this effect is easily faked.

metacognitive ability (Fleming and Lau, 2014). But they have rarely been themselves a focus of computational modeling. Broadly, this dissertation focuses on exploring the computational underpinnings of the subjective feeling of confidence, as measured through explicit confidence ratings. More specifically, we aim to test a normative model of confidence ratings as being a function of the posterior probability of being correct.

Some of the oldest models of confidence in binary categorization are based on signal detection theory (SDT) (Green and Swets, 1966). In signal detection theory, the observer has a noisy measurement that came from one of two categories. To determine their category choice, they compare their measurement to a criterion. SDT confidence models posit that the observer's confidence is the distance between the measurement and the criterion (Vickers, 1979).

SDT confidence models make an unusual prediction in situations where sensory noise is variable. In such models, an observer reports high confidence whenever the measurement falls above some criterion. Assuming that these criteria are fixed across noise conditions, on trials with high noise, measurements are increasingly likely to fall into the "high confidence" bin, even though average performance will be lower.[IV]

As an alternative to SDT models, several researchers have proposed a Bayesian alternative: confidence should be defined as the observer's posterior probability of being correct (Drugowitsch et al., 2014b; Hangya et al., 2016; Kepecs and Mainen, 2012; Meyniel et al., 2015; Pouget et al., 2016). Bayesian decision theory provides a general, normative, and often quantitatively accurate account of perceptual decisions in a wide variety of tasks in which an

[IV] Although we use this unusual prediction of the SDT models as a way to partly motivate a Bayesian model of confidence, we should note that Rahnev et al. (2011) do find evidence that humans report higher visibility for more noisy stimuli. They take this as evidence that humans use visibility criteria that are fixed across noise conditions. One caveat with their finding is that asking subjects to report visibility causes them to adopt a more conservative strategy than if asked to report confidence (Rausch and Zehetleitner, 2016); this may partly explain the results of Rahnev et al. (2011).

organism has noisy sensory input (Knill and Richards, 1996; Körding, 2007; Ma, 2012; Ma and Jazayeri, 2014). According to this theory, the Bayesian observer combines knowledge about the statistical structure of the world with the present sensory input to compute a posterior probability distribution over possible states of the world. In categorization tasks, a Bayesian model computes the log posterior odds of one category as the decision variable and makes a decision by comparing that decision variable to some criterion based on category base rate (i.e., prior) and the expected reward from either category. Computing that decision variable requires that an organism knows the noise associated with its sensory measurement. A Bayesian model of confidence is, conceptually, a simple extension to the Bayesian model of choice; it uses the Bayesian decision variable to determine confidence as well as choice.[V] In contrast to the above-mentioned SDT confidence models, Bayesian models always take measurement noise into account. The decision variable in a Bayesian model of confidence can be directly mapped onto the observer's posterior probability that the observer is correct.

Defining confidence as Bayesian would seem to allow neuroscientists to search for neural activity that appears to be "confidence" signals, confidence being thus defined (Kepecs and Mainen, 2012; Kepecs et al., 2008). This is a fine approach for many research questions. But although this definition of confidence has normative appeal, it may not be justified by evidence; there are no studies that convincingly show that Bayesian models provide better descriptions of confidence than other models. Here, we treat the notion that confidence is the posterior probability of being correct not as a definition, but as a hypothesis, which we

[V] All models in this dissertation, Bayesian or non-Bayesian, consider category choice and confidence report to be derived from the same decision variable. Because of this, our work does not directly address recent modeling efforts that treat category choice formation and confidence as emerging from distinct processes (Fleming and Daw, 2017; Moran et al., 2015; Pleskac and Busemeyer, 2010; van den Berg et al., 2016).

call the Bayesian confidence hypothesis (BCH).[VI]

The primary goal of this dissertation is to test the BCH, which requires that we not predefine confidence as Bayesian. Indeed, defining confidence as Bayesian appears to preclude a meaningful test of the BCH. We instead define confidence as "what humans report when you ask how confident they are." We use this definition because human subjects are able to easily report these naïve confidence ratings on a task, and because using this report as a definition allows us to determine the computations that underlie confidence. It has been proposed that such semantic and non-mathematical definitions of confidence are problematic (Kepecs and Mainen, 2012). However, for asking if human confidence reports are Bayesian, we see no other clear option. Surprisingly, no previous studies have asked whether naïve confidence ratings are Bayesian.

Recent research on whether confidence can be considered Bayesian has fallen into one of two broad approaches. The first approach is to derive qualitative patterns (i.e., "signatures") that should be visible in the data if the data were generated by a Bayesian observer (Hangya et al., 2016). Following the derivation of these signatures, one can plot behavioral data, look to confirm the presence of the signatures, and draw conclusions about what underlying model is likely to have produced the data (Navajas et al., 2017; Sanders et al., 2016). It would be hugely convenient if this approach were likely to yield scientific insight, because it would eliminate the need for painstaking quantitative work. But, in addition to other issues that we will discuss, this approach is ill-conceived, as they are neither necessary nor sufficient conditions for the BCH. The second category is to fit computational models to confidence ratings and compare the qualities of the fits. This approach is consistent with work that has frequently been done in the perceptual categorization literature, but rarely (Aitchison et al.,

[VI] This hypothesis is natural for categorization tasks, but might not be natural for other tasks. In an estimation task, for instance, it might be more natural to test whether confidence is a function of the posterior variance.

2015) in the confidence literature.

1.3 Dissertation outline

This dissertation will consist of a theoretical chapter making the case in favor of quantitative over qualitative techniques for testing the BCH, two experimental chapters, one exploratory chapter, and a conclusion.

In Chapter 2, we describe the general task setup that we will use and describe a technique that has been proposed for testing the BCH. Recently, authors have proposed gathering evidence in favor of the BCH by observing whether qualitative signatures of Bayesian confidence are present in data. We critically discuss this technique and conclude that quantitative model comparison is required.

In Chapter 3, we collect confidence reports from human subjects in binary categorization tasks. In an attempt to do a thorough test of the BCH, we fit and compare dozens of computational models. In this chapter, stimulus uncertainty is induced by manipulating external factors such as stimulus contrast.

Chapter 4 is an extension of the work in Chapter 3, except that stimulus uncertainty is induced by manipulating an internal factor: subjects' attention levels.

In Chapter 5, we describe an exploratory analysis of neural networks that may offer insight into why the brain might converge upon a heuristic solution to confidence. We intend this chapter to be a proof of concept, rather than a fully developed analysis.

Finally, in Chapter 6, we describe some caveats of our work and our thoughts on the future of confidence research.

Chapter 2

Limitations of proposed signatures of Bayesian confidence

2.1 Introduction

In recent years, some researchers have tested the Bayesian confidence hypothesis (BCH) by formally comparing Bayesian confidence models to other models (Aitchison et al., 2015). Although this is the most thorough method to test the BCH, it can be painstaking in practice. To avoid this approach, one could instead try to describe qualitative patterns that should theoretically emerge from Bayesian confidence and then look for those patterns in real data. Partly following this motivation, Hangya et al. (2016) propose signatures of the BCH, some of which have been observed in behavior (Kepecs et al., 2008; Lak et al., 2014; Sanders et al., 2016) and in neural activity (Kepecs et al., 2008; Komura et al., 2013).

These signatures are not unique to the Bayesian model; instead, they are expected under a number of other models (Kepecs and Mainen, 2012). This may be considered an advantage for a confidence researcher who is not interested in the precise algorithmic underpinnings of confidence. A researcher may observe these signatures in behavior, reasonably conclude that she has evidence that the organism is computing some form of confidence, and probe more deeply into, for instance, neural activity (Kepecs et al., 2008). In this manuscript, however, we consider the researcher concerned with understanding the algorithm used by an organism to compute confidence. For such a researcher, the fact that these signatures emerge from multiple models poses a problem: These signatures are not sufficient conditions for any particular model of confidence, including the Bayesian model. In other words, observation of these signatures does not constitute strong evidence in favor of any particular model. Because of this insufficiency, we view with skepticism any research that uses observation of these signatures as the basis for a claim that an organism uses a Bayesian form (Navajas et al., 2017), "statistical" form (Sanders et al., 2016), or any other specific form of confidence.

Although they do not claim that the signatures are sufficient conditions, Hangya et al. do claim that the signatures are necessary conditions for the BCH, i.e., that if confidence is Bayesian, these patterns will be present in behavior. If the signatures are necessary but not sufficient conditions for the BCH, observation of a single signature does not imply that the BCH is true; instead, one would need to observe several signatures in order to gain confidence in the nature of confidence.[I] However, we show that two of these signatures are not necessary conditions, reducing the overall value of the qualitative signature method for testing the BCH.

One signature is a mean confidence (i.e., the observer's estimated probability of being correct) of 0.75 for trials with neutral evidence. We show that, under the Bayesian model, this signature will only be observed when noise is very low and stimulus distributions do not overlap.

[I] Restating this logic in probabilistic terms: A signature being a necessary condition for the BCH implies that p(signature observed | BCH is true) $= 1$. A signature being an insufficient condition implies that p(signature observed | BCH is false) > 0. By Bayes' rule, for signatures that are both necessary and insufficient, p(BCH is true | signature(s) observed) will increase with the observation of each signature but will never reach 1.

Another signature is that, as stimulus magnitude increases, mean confidence increases on correct trials but decreases on incorrect trials. Here, we show that under the Bayesian model, this signature breaks down when noise is low and stimulus distributions are Gaussian. We also explain and resolve a recent discrepancy in the literature that is related to an alternative formulation of this signature (Navajas et al., 2017).

2.2 Binary categorization task

We restrict ourselves to the following, widely used, family of binary perceptual categorization tasks (Green and Swets, 1966). On each trial, a category $C \in \{-1, 1\}$ is randomly drawn with equal probability. Each category corresponds to a stimulus distribution $p(s \mid C)$, where s may specify the value of many possible kinds of stimuli (e.g., an odor mixture (Kepecs et al., 2008), the net motion energy of a random dot kinematogram (Kiani and Shadlen, 2009; Newsome et al., 1989), the orientation of a Gabor (Chapters 3 and 4 of this dissertation; Qamar et al., 2013), or the mean orientation of a series of Gabors (Navajas et al., 2017)). The stimulus distributions are mirrored across $s = 0$, i.e., $p(s \mid C = -1) = p(-s \mid C = 1)$. We assume that the observer has full knowledge of these distributions. A stimulus s is drawn from the chosen stimulus distribution and presented to the observer. The observer does not have direct access to the value of s; instead, they take a noisy measurement x, drawn from the distribution $p(x \mid s) = \mathcal{N}(x; s, \sigma)$, which denotes a Gaussian distribution over x with mean s and standard deviation σ (Figure 2.1).[II] The above description applies for the tasks used in this dissertation, except that, in Task B (Chapters 3 to 5), the stimulus distributions are not mirrored across $s = 0$.

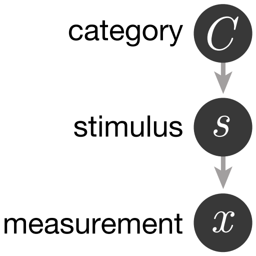

Figure 2.1 Generative model of the task.

If the observer's choice behavior is Bayes-optimal (i.e., minimizes expected loss which, in a task where each category has equal reward, is equivalent to maximizing accuracy), they compute the posterior probability of each category by marginalizing over all possible values of s: $p(C \mid x, \sigma) = \int p(C \mid s) p(s \mid x, \sigma) \, \mathrm{d}s$. They then make a category choice $\hat{C}$ by choosing the category with the highest posterior: $\hat{C} = \operatorname{argmax}_C p(C \mid x, \sigma)$. For mirrored stimulus distributions, that amounts to choosing $\hat{C} = 1$ when $x > 0$, and $\hat{C} = -1$ otherwise.

Furthermore, if the observer's confidence behavior is Bayesian, then it will be some function of the posterior probability of the chosen category. This probability is

[II] Because some of our notation relates to that used in Hangya et al. (2016), we provide this table to enable easier comparison between the two papers. In some cases, the variables are not exactly identical: the terms in Hangya et al. may be more general. This does not affect the validity of our claims. For consistency, we always describe their work using our notation.

This dissertation	**Hangya et al. (2016)**
true category C	not used
stimulus s	evidence d
stimulus magnitude $\lvert s \rvert$	discriminability Δ
measurement x	percept $\hat{d}$
choice $\hat{C}$	choice ϑ
confidence $p(C = \hat{C} \mid x, \sigma) = \mathrm{conf}(x, \sigma)$	confidence $c = \xi(\hat{d}, \vartheta)$

$p(C = \hat{C} \mid x, \sigma) = \max_C p(C \mid x, \sigma)$. Because it is a deterministic function of x and σ, we will refer to it as $\text{conf}(x, \sigma)$.[III]

2.3 Derivation of Bayesian confidence

We will now derive $\text{conf}(x, \sigma)$ for all stimulus distributions used in this chapter (other chapters will use only Gaussian stimulus distributions). As described in Section 2.2, if an observer's confidence behavior is Bayesian, it is a function of the posterior probability of the most probable category. By Bayes' rule,

$$\begin{aligned}\text{conf}(x, \sigma) &= \max_C p(C \mid x) \\ &= \max_C \frac{p(x \mid C)p(C)}{\sum_C p(x \mid C)p(C)} \\ &= \max_C \frac{p(x \mid C)}{\sum_C p(x \mid C)}. \end{aligned} \tag{2.1}$$

In the last step, we eliminated the prior because each category is equally likely (i.e., $p(C = 1) = p(C = -1)$) and we assume that the observer knows this. We now derive the task-specific likelihood functions $p(x \mid C)$ used in our simulations. The observer does not know the true stimulus value s, but does know that the measurement is drawn from a Gaussian distribution with a mean of s and s.d. σ. Using this knowledge, the optimal observer marginalizes over s by convolving the stimulus distributions with their noise distribution:

$$\begin{aligned}p(x \mid C) &= \int p(x \mid s) p(s \mid C) \, \mathrm{d}s \\ &= \int \mathcal{N}(x; s, \sigma) p(s \mid C) \, \mathrm{d}s. \end{aligned} \tag{2.2}$$

[III] Note that our assumption that confidence and category choice are deterministic functions of x amounts to an assumption that there is no noise at the action (i.e., reporting) stage.

For uniform category distributions, we plug $p(s \mid C) = \mathcal{U}(s; a, b)$, which denotes a continuous uniform distribution over s between a and b, into Equation (2.6) and simplify:

$$\begin{aligned} p_{\mathrm{U}}(x \mid C) &= \int \mathcal{N}(x; s, \sigma)\mathcal{U}(s; a, b)\,\mathrm{d}s \\ &= \frac{1}{b-a}\int_a^b \mathcal{N}(x; s, \sigma)\,\mathrm{d}s \\ &= \frac{1}{\sigma(b-a)}\left(\Phi(b-x) - \Phi(a-x)\right), \end{aligned} \tag{2.3}$$

where Φ is the cumulative distribution function of the standard normal distribution. For Gaussian category distributions, we plug $p(s \mid C) = \mathcal{N}(s; \mu_C, \sigma_C)$ into Equation (2.6) and simplify:

$$\begin{aligned} p_{\mathrm{G}}(x \mid C) &= \int \mathcal{N}(x; s, \sigma)\mathcal{N}(s; \mu_C, \sigma_C)\,\mathrm{d}s \\ &= \mathcal{N}\left(x; \mu_C, \sqrt{\sigma^2 + \sigma_C^2}\right), \end{aligned} \tag{2.4}$$

using $\sigma_C = 0$ if stimuli from a given category always take on the same value μ_C.

Finally, plug the task-appropriate likelihood function (Equation (2.7) or Equation (2.8)) into Equation (2.5).

2.4 0.75 signature: Mean Bayesian confidence is 0.75 for neutral evidence trials.

Hangya et al. (2016) propose a signature concerning neutral evidence trials, those in which the stimulus s is equal to 0 (i.e., there is equal evidence for each category), and observer performance is at chance. Bayesian confidence on each individual trial will always be at least 0.5 (assuming that measurement noise is nonzero). One can intuitively understand why this is: in binary categorization, if the posterior probability of one option is less than 0.5, the observer makes the other choice, which has a posterior probability above 0.5. Therefore, all trials have confidence of at least 0.5, and mean confidence at any value of s is also greater than 0.5. Hangya et al. go beyond these results and provide a proof that, under some assumptions, mean Bayesian confidence on neutral evidence trials is *exactly* 0.75.[IV] We refer to this prediction as the 0.75 signature, and we show that it is not always expected under a normative Bayesian model.

2.4.1 The 0.75 signature is not a necessary condition for Bayesian confidence

To determine the conditions under which the 0.75 signature is expected under the Bayesian model, we used Monte Carlo simulation with the following procedure. For a range of measurement noise levels σ, we drew measurements x from $\mathcal{N}(x; s = 0, \sigma)$. Using the

[IV] The proof of the 0.75 signature depends on a lemma proved by Hangya et al. (2016): *Integrating the product of the probability density function f and the distribution function F of any probability distribution symmetric to zero over the positive half-line results in 3/8:*

$$\int_0^\infty f(t)F(t)\mathrm{d}t = \frac{3}{8}.$$

There is a shorter proof of the lemma, which is as follows. Use integration by parts, and that $f(t) = F'(t)$ by definition:

$$\begin{aligned}\int_0^\infty f(t)F(t)\,\mathrm{d}t &= F(\infty)F(\infty) - F(0)F(0) - \int_0^\infty f(t)F(t)\,\mathrm{d}t\\ 2\int_0^\infty f(t)F(t)\,\mathrm{d}t &= F(\infty)F(\infty) - F(0)F(0).\end{aligned}$$

Because F is a cumulative distribution function of a probability distribution symmetric across zero, $F(\infty) = 1$ and $F(0) = \frac{1}{2}$:

$$\begin{aligned}2\int_0^\infty f(t)F(t)\,\mathrm{d}t &= 1 - \frac{1}{4}\\ \int_0^\infty f(t)F(t)\,\mathrm{d}t &= \frac{3}{8}.\end{aligned}$$

function $\text{conf}(x, \sigma)$ that the observer would use if they believed stimuli were being drawn from category-conditioned stimulus distributions $p(s \mid C)$ (rather than all s being zero), we computed Bayesian confidence for each measurement. We then took the mean confidence, equal to $\mathbb{E}_{x|s=0}\left[\text{conf}(x, \sigma)\right]$.

The 0.75 signature only holds if the s.d. of the noise is very low relative to the range of the stimulus distribution. Additionally, the observer must believe that the category-conditioned stimulus distributions are non-overlapping (Figure 2.2a, dotted line). If the observer believes that the category-conditioned stimulus distributions overlap by even a small amount, mean confidence on neutral evidence trials drops to 0.5. Therefore, in an experiment with overlapping stimulus distributions, one should not expect an optimal observer to produce the 0.75 signature. In experiments with non-overlapping distributions, an observer's false belief about the distributions might also cause them to not produce the 0.75 signature. We use the example of overlapping uniform stimulus distributions (Figure 2.2a, solid lines) to demonstrate the fragility of this signature, although such distributions are not common in the literature. Overlapping Gaussian stimulus distributions (Figure 2.2b), however, are relatively common in the perceptual categorization literature (Ashby and Gott, 1988; Green and Swets, 1966; Norton et al., 2017; Qamar et al., 2013) and arguably more naturalistic (Maddox, 2002). Because the 0.75 signature requires both low measurement noise and non-overlapping stimulus distributions, mean 0.75 confidence at neutral evidence trials is not a necessary condition for Bayesian confidence.

Additionally, the 0.75 signature is only relevant in experiments where subjects are specifically asked to report confidence in the form of a perceived probability of being correct (or are incentivized to do so through a scoring rule (Brier, 1950; Gneiting and Raftery, 2007; Massoni et al., 2014), although in this case it has been argued (Ma and Jazayeri, 2014) that any Bayesian behavior might simply be a learned mapping). In other words, in an experiment where subjects are asked to report confidence on a 1 through 5 scale, a mean confidence of 3 only corresponds to 0.75 if one makes the a priori assumption that there is a linear mapping between rating and perceived probability of being correct (Sanders et al., 2016).

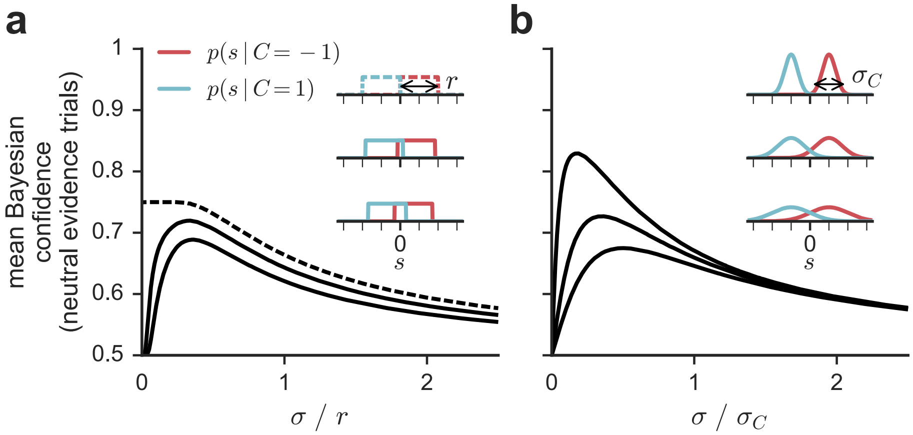

Figure 2.2 The 0.75 signature is not a necessary condition for Bayesian confidence. The y-axis indicates mean Bayesian confidence on trials for which $s = 0$. Each inset corresponds to a line, in the same top-to-bottom order. Dotted and solid lines indicate, respectively, non-overlapping and overlapping categories. For each value of σ, 50,000 trials were simulated. (**a**) Trials were simulated using uniform stimulus distributions defined by $p(s \mid C = 1) = \mathcal{U}(s; a, b)$, with $b - a = r = 2$. When the stimulus categories are non-overlapping (i.e., with $a = 0$ and $b = 2$, top inset), the 0.75 signature can be observed at zero measurement noise (dotted black line). However, mean Bayesian confidence decreases as a function of measurement noise. Additionally, when the distributions overlap slightly (bottom two insets), the 0.75 signature will not be observed (solid black lines). (**b**) Moreover, when the stimulus categories are Gaussian distributions defined by $p(s \mid C = 1) = \mathcal{N}(s; \mu_C = 1, \sigma_C)$, the 0.75 signature will not be observed at any σ_C or measurement noise level σ. One can intuitively understand why mean confidence is 0.5 for overlapping categories at very low measurement noise and increases with measurement noise. At very low measurement noise, the observer makes measurements that are very close to zero, which the observer "knows" are associated with a low probability of being correct. However, as noise increases, the observer starts to make measurements that have higher magnitude, leading the observer to believe that they have a higher probability of being correct.

2.4.1.1 Relevant assumptions in Hangya et al.

Hangya et al. describe an assumption that is critical for the 0.75 signature: each category-conditioned stimulus distribution is a continuous uniform distribution. However, the 0.75

signature depends on two additional assumptions that they make implicitly.

Their proof depends on confidence for one category being equal to $p(s > 0 \mid x, \sigma)$ (p. 1852). This equality further depends on their implicit assumption both of non-overlapping categories and of negligible measurement noise; these assumptions are equivalent to only considering the leftmost point of the solid line in Figure 2.2a. To understand why, we derive their definition of confidence as $p(s > 0 \mid x, \sigma)$.

Without loss of generality, we look at trials with choice $\hat{C} = 1$. First, Hangya et al. make the assumption that the categories are non-overlapping uniforms (i.e., $p(s \mid C = 1) = \mathcal{U}(s; 0, b)$). This allows them to write (Section 2.4.2):

$$\begin{aligned}\mathrm{conf}_{\hat{C}=1}(x, \sigma) &= \frac{p(x \mid C = 1)}{p(x \mid C = 1) + p(x \mid C = -1)} \\ &= \frac{\int_0^b p(x \mid s, \sigma)\, \mathrm{d}s}{\int_0^b p(x \mid s, \sigma)\, \mathrm{d}s + \int_{-b}^0 p(x \mid s, \sigma)\, \mathrm{d}s}\end{aligned}$$

Second, they make the assumption that b is very large relative to measurement noise σ. This allows them to write:

$$\begin{aligned}\mathrm{conf}_{\hat{C}=1}(x, \sigma) &\approx \frac{\int_0^\infty p(x \mid s, \sigma)\, \mathrm{d}s}{\int_0^\infty p(x \mid s, \sigma)\, \mathrm{d}s + \int_{-\infty}^0 p(x \mid s, \sigma)\, \mathrm{d}s} \\ &\approx \int_0^\infty p(x \mid s, \sigma)\, \mathrm{d}s \\ &\approx p(s > 0 \mid x, \sigma).\end{aligned}$$

If the stimulus distributions overlap by even a small amount or if measurement noise is non-negligible, confidence cannot be written as $p(s > 0 \mid x, \sigma)$, and the proof of the 0.75 signature breaks down.

2.4.2 The 0.75 signature is not a sufficient condition for Bayesian confidence

We have shown that the 0.75 signature is not a necessary condition for Bayesian confidence, but is it a sufficient condition? It is possible to show that a signature is a sufficient condition if it is not possible to observe it under any other model. However, one could put forward a trivial model that always produces exactly midrange confidence on each trial, regardless of the measurement. Therefore, the 0.75 signature is not a sufficient condition.

We will now derive $\mathrm{conf}(x, \sigma)$ for all stimulus distributions used in this chapter (other chapters will use only Gaussian stimulus distributions). As described in Section 2.2, if an observer's confidence behavior is Bayesian, it is a function of the posterior probability of the most probable category. By Bayes' rule,

$$\begin{aligned}\mathrm{conf}(x, \sigma) &= \max_C p(C \mid x)\\ &= \max_C \frac{p(x \mid C)p(C)}{\sum_C p(x \mid C)p(C)}\\ &= \max_C \frac{p(x \mid C)}{\sum_C p(x \mid C)}.\end{aligned} \tag{2.5}$$

In the last step, we eliminated the prior because each category is equally likely (i.e., $p(C = 1) = p(C = -1)$) and we assume that the observer knows this. We now derive the task-specific likelihood functions $p(x \mid C)$ used in our simulations. The observer does not know the true stimulus value s, but does know that the measurement is drawn from a Gaussian distribution with a mean of s and s.d. σ. Using this knowledge, the optimal observer marginalizes over s by convolving the stimulus distributions with their noise distribution:

$$\begin{aligned}p(x \mid C) &= \int p(x \mid s)p(s \mid C)\,\mathrm{d}s\\ &= \int \mathcal{N}(x; s, \sigma)p(s \mid C)\,\mathrm{d}s.\end{aligned} \tag{2.6}$$

For uniform category distributions, we plug $p(s \mid C) = \mathcal{U}(s; a, b)$, which denotes a continuous uniform distribution over s between a and b, into Equation (2.6) and simplify:

$$\begin{aligned} p_{\mathrm{U}}(x \mid C) &= \int \mathcal{N}(x; s, \sigma)\mathcal{U}(s; a, b)\,\mathrm{d}s \\ &= \frac{1}{b-a}\int_a^b \mathcal{N}(x; s, \sigma)\,\mathrm{d}s \\ &= \frac{1}{\sigma(b-a)}\left(\Phi(b-x) - \Phi(a-x)\right), \end{aligned} \tag{2.7}$$

where Φ is the cumulative distribution function of the standard normal distribution. For Gaussian category distributions, we plug $p(s \mid C) = \mathcal{N}(s; \mu_C, \sigma_C)$ into Equation (2.6) and simplify:

$$\begin{aligned} p_{\mathrm{G}}(x \mid C) &= \int \mathcal{N}(x; s, \sigma)\mathcal{N}(s; \mu_C, \sigma_C)\,\mathrm{d}s \\ &= \mathcal{N}\left(x; \mu_C, \sqrt{\sigma^2 + \sigma_C^2}\right), \end{aligned} \tag{2.8}$$

using $\sigma_C = 0$ if stimuli from a given category always take on the same value μ_C.

Finally, plug the task-appropriate likelihood function (Equation (2.7) or Equation (2.8)) into Equation (2.5).

2.5 Divergence signature #1: As stimulus magnitude increases, mean confidence increases on correct trials but decreases on incorrect trials

Hangya et al. (2016) propose the following pattern as a signature of Bayesian confidence: On correctly categorized trials, mean confidence is an increasing function of stimulus magnitude (here, $|s|$), but on incorrect trials, it is a decreasing function (Figure 2.3a). We refer to this pattern as the divergence signature.[V] The signature is present in Bayesian confidence when category-conditioned stimulus distributions are uniform, in both high- and low-noise regimes (Figure 2.3a,b). The intuition for why this pattern may occur is as follows. On correct trials, as stimulus magnitude increases, the mean magnitude of the measurement x increases. Because measurement magnitude is monotonically related to Bayesian confidence, this increases mean confidence. However, on incorrect trials (in which x and s have opposite signs), the mean magnitude of the measurement decreases (Figure 2.5a), which in turn decreases mean confidence (Figure 2.5b,c).

The divergence signature has been observed in some behavioral experiments (Kepecs et al., 2008; Komura et al., 2013; Lak et al., 2014; Sanders et al., 2016). However, we demonstrate that, as with the 0.75 signature the divergence signature is not always expected under a normative Bayesian model.[VI] Therefore, the appearance of the signature in these papers should not be taken to mean that it should be generally expected.

2.5.1 Divergence signature #1 is not a necessary condition for Bayesian confidence

To determine the conditions under which the divergence signature is expected under the Bayesian model, we used Monte Carlo simulation with the following procedure. We generated

[V] Kepecs and Mainen (2012), Insabato et al. (2016), and Fleming and Daw (2017) call it the (folded) "X-pattern."

[VI] Our finding is distinct from that of Insabato et al. (2016), who show that the divergence signature would not be predicted under a non-Bayesian model in which the observer uses two measurements on each trial. Our analyses only concern Bayesian models in which the observer has a single measurement on each trial.

Our finding is also distinct from that of Fleming and Daw (2017), who show that the divergence signature would not be predicted if the experimenter could plot confidence as a function of the internal measurement x. Our analyses only concern confidence as a function of stimulus magnitude $|s|$ which, unlike x, is known by the experimenter.

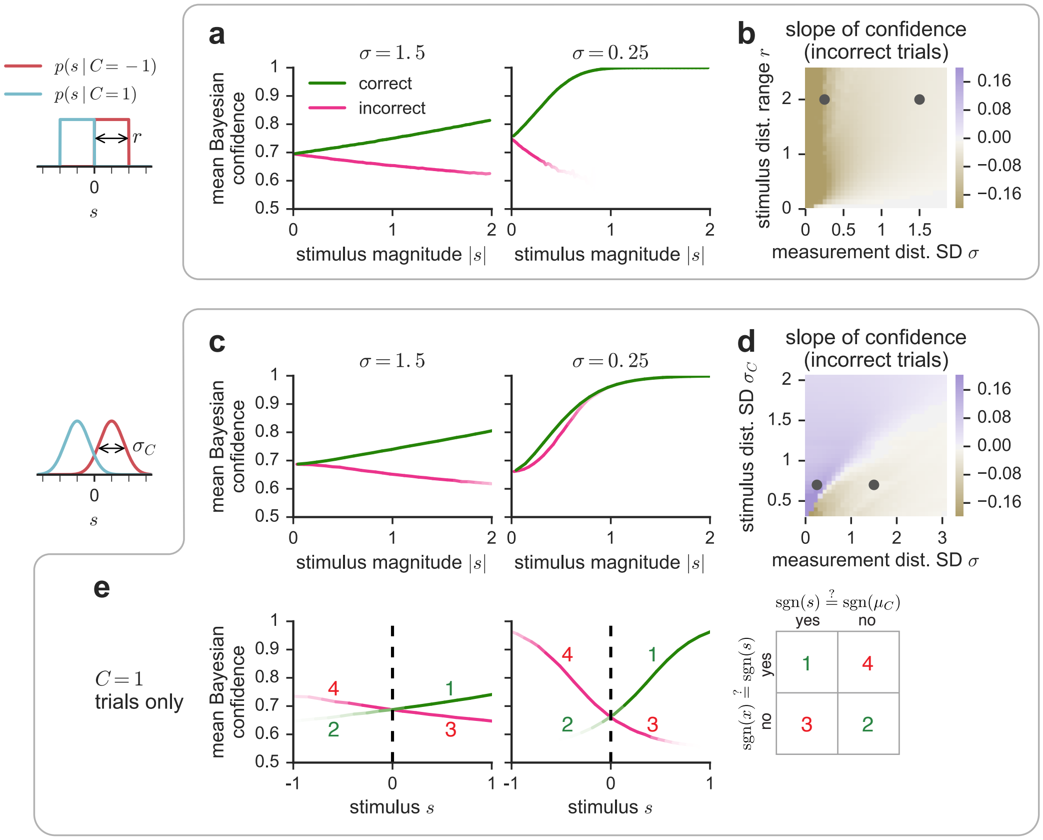

Figure 2.3 The divergence signature is not a necessary condition for Bayesian confidence. For two stimulus distribution types, we simulated 2 million trials. (**a**) With uniform stimulus distributions defined by $p(s \mid C = 1) = \mathcal{U}(s; 0, 2)$, the divergence signature is predicted under both high- and low-noise regimes. The fadedness of the line indicates conditions for which there are few trials. (**b**) Heatmap indicates the slope of the pink lines in **a**. At all values of σ and distribution range, the slope is negative. Slopes were obtained by generating binned mean confidence values as in **a** and fitting a line to those values. Black markers indicate the parameters used in **a**, with left dot corresponding to right plot and conversely. (**c**) With Gaussian stimulus distributions defined by $p(s \mid C = 1) = \mathcal{N}(s; 1, \sigma_C = 0.7)$, the divergence signature appears only when measurement noise is high, i.e., when $\sigma \lesssim 0.6$. (**d**) As in **b** but for Gaussian distributions with means of ± 1. Under some values of σ and σ_C, the slope is positive, indicating that the divergence signature is not a necessary condition for Bayesian confidence. (**e**) Visual explanation for why, under Gaussian stimulus distributions, the divergence signature appears only at relatively high σ values. Plots represent the same data as in **c**, but over s instead of $|s|$. For clarity, we only use trials drawn from category $C = 1$; the argument is unaffected. Incorrect trials fall into two categories: on trials in which s is positive but x is negative due to noise, confidence goes down as $|s|$ increases (branch 3); on trials in which s and x are both negative, confidence increases with $|s|$ (branch 4). At high levels of noise, branch 3 has more trials than branch 4, and dominates the averaging that occurs when plotting trials from both categories over $|s|$. At low levels of noise, branch 4 instead dominates, and the divergence signature disappears. Note that, for non-overlapping distributions (e.g., those in **a**,**b**), there are no trials in which s has a different sign than the stimulus distribution mean, so branches 2 and 4 do not exist, and the divergence signature is always present.

stimuli s, drawn with equal probability from stimulus distributions $p(s \mid C = -1)$ and $p(s \mid C = 1)$. We generated noisy measurements x from these stimuli, using measurement noise levels σ. We generated observer choices from these measurements, using the optimal decision rule $x > 0 \Rightarrow \hat{C} = 1$, and we computed Bayesian confidence for every trial.

When stimulus distributions are Gaussian and measurement noise is low relative to stimulus distribution width, the divergence signature is not expected (Figure 2.3c,d). To understand why this is, imagine an optimal observer with zero measurement noise. In tasks with overlapping categories, even this observer cannot achieve perfect performance; for a given category with a positive mean, there are stimuli that have a negative value, resulting in an incorrect choice. For such stimuli, confidence *increases* with stimulus magnitude. At relatively low noise levels, these stimuli represent the majority of all incorrect trials for the category (Figure 2.3e). This effect causes the divergence signature to disappear when averaging over trials drawn from both categories. Because of this, the divergence signature is not a necessary condition for Bayesian confidence. Note that an experimenter could avoid this issue by plotting confidence as a function of signed stimulus value s and by not averaging over both categories, which would produce plots such as Figure 2.3e.

2.5.1.1 Relevant assumption in Hangya et al.

We have shown that the applicability of the divergence signature may be limited to particular cases. By contrast, the proof in Hangya et al. suggests that it is quite general. We can resolve this paradox by making explicit the assumptions hidden in the proof. They assume that, "for incorrect choices. . . with increasing evidence discriminability, the relative frequency of low-confidence percepts increases while the relative frequency of high-confidence percepts decreases" (p. 1847).[VII] This assumption is violated in the case of overlapping Gaussian stimulus distributions: for some incorrect choices (branch 4 of Figure 2.3e), as s becomes more discriminable (i.e., very negative), the frequency of *high*-confidence reports increases. At low levels of measurement noise, this causes the divergence signature to disappear.

2.5.2 Divergence signature #1 is not a sufficient condition for Bayesian confidence

It has been previously noted that the signature is expected under a number of non-Bayesian models (Fleming and Daw, 2017; Insabato et al., 2016; Kepecs and Mainen, 2012). Here, we describe an additional non-Bayesian model, one in which confidence is a function only of $|x|$, the magnitude of the measurement.[VIII] In the general family of binary categorization tasks described in Section 2.2, the confidence of this model is monotonically related to the confidence of the Bayesian model $\text{conf}(x, \sigma)$. Thus, when the divergence signature is predicted by the Bayesian model, it is also predicted by this measurement model. Therefore, the divergence signature is not a sufficient condition for Bayesian confidence.

[VII] Their original assumption actually reads, "for any given confidence c, the relative frequency of percepts mapping to c by ξ changes monotonically with evidence discriminability for any fixed choice." In our terminology, this is equivalent to saying that, as $|s|$ increases, the frequency of reporting any particular level of confidence changes monotonically. This is not correct even in the case of uniform stimulus distributions; for example, at low noise, as discriminability increases, the frequency of medium-confidence reports will increase and then decrease. Their restatement of this assumption specifically for incorrect choices, which we cite in the main text, is correct for non-overlapping stimulus distributions. Because they restate the assumption correctly, their following argument holds except under the scenario described in the main text.

[VIII] In the Bayesian model, observers use their knowledge of their uncertainty. In this alternative standard signal detection theoretical model (Green and Swets, 1966; Kepecs and Mainen, 2012), observers ignore uncertainty, making confidence only a function of the distance between the measurement and the decision bound. Previous studies have referred to similar models as Difference (Aitchison et al., 2015) or Fixed (Qamar et al., 2013). In Chapters 3 to 5, we will also call this model Fixed.

2.6 Divergence signature #2: As measurement noise decreases, mean confidence increases on correct trials but decreases on incorrect trials

An alternative version of the divergence signature has emerged in the literature. Navajas et al. (2017) conduct an experiment in which they present, on each trial, a series of oriented Gabors with orientations pseudorandomly drawn from uniform distributions with different variances. They then ask subjects to judge whether the mean orientation is left or right of vertical and to provide a confidence report. They plot confidence as a function of correctness and orientation distribution variance, expecting that, if confidence were Bayesian, their data would look like Figure 2.3a. Contrary to their expectations, they observe no such divergence (Figure 2.4a). However, instead of plotting *stimulus magnitude*, which produces divergence signature #1 (Section 2.5), they plot *measurement noise*[IX] on the x-axis (Figure 2.4a), in effect proposing a divergence signature distinct from the one described in Section 2.5. We will refer to this as divergence signature #2: as measurement noise decreases, mean confidence increases on correct trials but decreases on incorrect trials. We find no evidence that this signature is expected under the Bayesian confidence model, resolving the seemingly unexpected result in Navajas et al.

[IX] However, because the orientations were drawn such that the mean orientation of each set was the same for all trials in a category, there was no variance over the stimulus variable of interest (the per-trial mean) within categories. Therefore, what they describe as stimulus variance factors into a Bayesian model of confidence (and into their non-Bayesian decision model) only by changing measurement noise. Additionally, because there is no variance over stimulus magnitude within categories, they are unable to determine whether divergence signature #1 is present in their data.

2.6.1 Divergence signature #2 is not expected under Bayesian confidence

To determine whether divergence signature #2 is expected under the Bayesian model, we used Monte Carlo simulation with the following procedure. We generated stimuli with $s = \pm 1$, corresponding to $C = \pm 1$.[X] For a range of measurement noise levels σ, we drew noisy measurements x from $\mathcal{N}(x; s, \sigma)$. We generated observer choices from these measurements, using the optimal decision rule $x > 0 \Rightarrow \hat{C} = 1$. We computed Bayesian confidence for every trial.

As measurement noise decreases, mean confidence increases for both correct and incorrect trials (Figure 2.4b). This pattern also holds when the category-conditioned stimulus distributions are uniform or Gaussian, and if one plots a measure of stimulus distribution variance on the x-axis (either uniform distribution range r or Gaussian distribution s.d. σ_C). This indicates that the signature is not expected under the BCH.

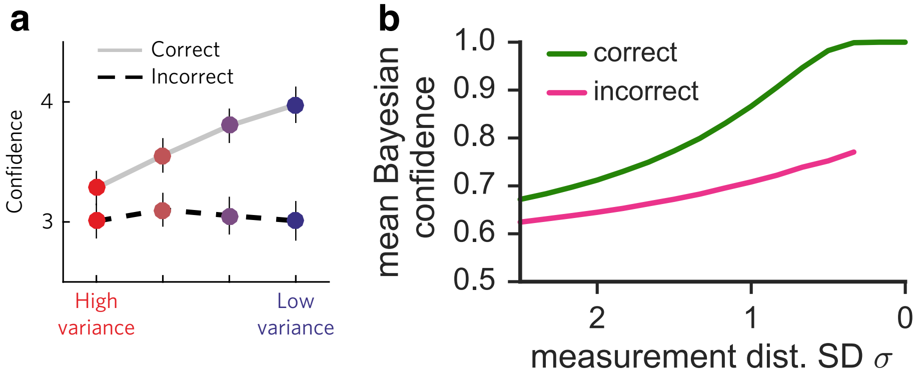

Figure 2.4 Divergence signature #2 is not present either in the Navajas et al. data or in the prediction of the Bayesian model. (**a**) Average confidence in a binary perceptual categorization task, reproduced with permission from Navajas et al. (2017). (**b**) Mean Bayesian confidence as a function of measurement noise is not expected to show opposite trends when conditioned on correctness. At each value of σ, 50,000 stimuli were stimulated, with $s = \pm 1$.

[X] This corresponds to Navajas et al. (2017), as described in Section 2.6.

2.6.1.1 Related text in Hangya et al.

It is quite understandable that Navajas et al. took measurement noise as their definition of evidence discriminability; Hangya et al. explicitly allow it in their description of the divergence signature: "any monotonically increasing function of expected outcome [i.e., accuracy]. . . can serve as evidence discriminability" (p. 1847). Measurement noise (or, in keeping strictly with Hangya et al.'s definition, measurement precision) is indeed monotonically related to accuracy. However, the divergence signature requires an additional assumption: "for incorrect choices. . . with increasing evidence discriminability, the relative frequency of low-confidence percepts increases while the relative frequency of high-confidence percepts decreases" (p. 1847; see also, Section 2.5.1.1). Simulation shows that this assumption is violated when measurement noise is used as the definition of evidence discriminability.

2.6.1.2 Why the intuition for divergence signature #1 does not predict divergence signature #2

We have shown that, although divergence signature #1 is not completely general, it is expected under the Bayesian model in some cases (Figure 2.3a). By contrast, there is no indication that divergence signature #2 is ever expected. This may be surprising, because the intuition for divergence signature #1 might seem to apply equally to divergence signature #2. However, the effect of measurement noise on mean confidence is different than the effect of stimulus magnitude because measurement noise, unlike stimulus magnitude, affects the mapping from measurement to confidence on a single trial.

Mean Bayesian confidence is a function of two factors: confidence on a single trial and the probability of the corresponding measurement.

$$\mathop{\mathbb{E}}_{x}\left[\mathrm{conf}(x,\sigma)\right] = \int \mathrm{conf}(x,\sigma) p(x \mid s,\sigma)\,\mathrm{d}x$$

The intuition for divergence signature #1 is as follows: as stimulus magnitude $|s|$ increases, the measurement distribution $p(x \mid s, \sigma)$ shifts, and the mean measurement magnitude on incorrect trials decreases (Figure 2.5a). One might expect this intuition to also result in divergence signature #2, since the effect of decreased measurement noise σ on $p(x \mid s, \sigma)$ also results in a decreased measurement magnitude on incorrect trials (Figure 2.5d). However, σ additionally affects $\text{conf}(x, \sigma)$, the per-trial, deterministic mapping from measurement and noise level to Bayesian confidence (Figure 2.5e), whereas stimulus magnitude does not (Figure 2.5b). Therefore, when σ is variable, the resulting effect on the measurement distribution is insufficient for describing the pattern of mean confidence on incorrect trials, requiring simulation. We simulated experiments as described in Section 2.5.1, and demonstrate why stimulus magnitude and measurement noise have different effects on mean confidence on incorrect trials (Figure 2.5).

2.6.2 Use of divergence signature #2 in Navajas et al.

Navajas et al. motivate their findings by first building a *non-Bayesian* model of confidence[XI] that does predict divergence signature #2, i.e., that, as measurement noise decreases, mean confidence decreases on incorrect trials. They then fail to observe the signature in their averaged data (Figure 2.4a), observing instead that confidence is constant on incorrect trials. Some subjects (e.g., subject 16 in their Figure 3), however, do show the signature. This leaves them with a puzzle—what model can describe the data? To answer this, they modify their model to incorporate Fisher information, which increases as measurement noise decreases.

[XI] In their model, which they label "normative," the observer continually updates a weighted average of each stimulus with the previous average. This model is not equivalent to (nor a supermodel of) the optimal model, which keeps a running sum of stimuli, dividing by N for each stimulus or at the end of the trial. They motivate their non-Bayesian model by the observation that recent samples have a relatively higher influence on subject decisions, but do not show fits of a fully Bayesian model to their data.

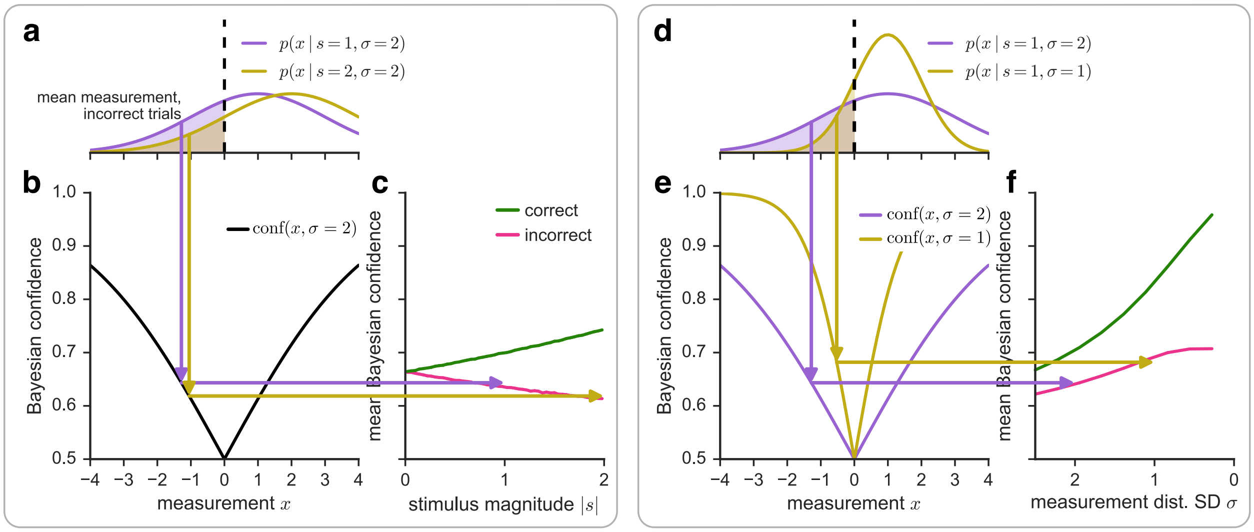

Figure 2.5 Explanation for why divergence signature #1 is sometimes expected, but why divergence signature #2 might not ever be expected. Although increased stimulus magnitude and decreased measurement noise both cause the mean measurement magnitude to decrease on incorrect trials, they have different effects on mean confidence. At each value of σ, 2 million stimuli were simulated, using uniform stimulus distributions defined by $p(s \mid C = 1) = \mathcal{U}(s; 0, 2)$ (the case of Figure 2.3a). (**a**) As described previously (Drugowitsch, 2016; Hangya et al., 2016; Kepecs et al., 2008), an increase in stimulus magnitude causes the mean measurement magnitude to decrease on incorrect trials. (**b**) Measurements are mapped onto confidence values using the deterministic function $\mathrm{conf}(x, \sigma)$, which is equivalent to the posterior probability that the choice is correct (Section 2.2). (**c**) This mapping results in divergence signature #1, a decrease in mean confidence on incorrect trials. Arrows do not align precisely with the simulated mean, because the confidence of the mean measurement is not exactly equal to the mean confidence. (**d**) A decrease in measurement noise also causes the mean measurement magnitude to decrease on incorrect trials. (**e**) Because the mapping from measurement to confidence $\mathrm{conf}(x, \sigma)$ is dependent on σ, measurements from the less noisy distribution have higher confidence. (**f**) Because the confidence mapping is dependent on σ, divergence signature #2 is not necessarily expected under Bayesian confidence.

This post-hoc model is able to "bend" the confidence curve upward as measurement noise decreases, producing curves that more closely resemble their data.

The main shortcoming of this argument is that a Bayesian model of confidence would not actually predict divergence signature #2, as we have shown above. Indeed, their averaged data more closely resembles the prediction of the Bayesian model (Figure 2.4b) than that of their non-Bayesian model without Fisher information (their Figure 2b). Therefore, the absence of the signature in their averaged data does not suggest anything beyond a Bayesian model; it is possible that the Bayesian model would provide a good fit to most of their subjects. If the model provided a poor fit to subjects that do show divergence signature #2, a post-hoc model would have to incorporate some other mechanism that could "bend" the confidence curve *downward*, which would not be Fisher information.

2.7 Other signatures

A third signature in Hangya et al. (2016) that we do not discuss here (that confidence equals accuracy), is like the 0.75 signature in that it either requires explicit reports of perceived probability of being correct, or the experimenter to choose a mapping between rating and perceived probability of being correct (Section 2.4.1). For any monotonic relationship between accuracy and confidence, it is likely that there is some mapping that equates the two, in which case the signature would not be a sufficient condition for the BCH.

A fourth signature (that confidence allows a better prediction of accuracy than stimulus magnitude alone) is, like divergence signature #1, also predicted by the measurement model (Section 2.5.2) and is therefore also not a sufficient condition for the BCH.

2.8 Discussion

We have demonstrated that, even in the relatively restricted class of binary categorization tasks that we consider here (Section 2.2), some signatures are neither necessary nor sufficient conditions for the BCH. Specifically, the 0.75 signature is only expected under non-overlapping stimulus distributions. Additionally, despite claims that divergence signature #1 is "robust to different stimulus distributions," (Kepecs and Mainen, 2012) it is only expected under non-overlapping stimulus distributions or under Gaussian stimulus distributions with high measurement noise. Because of their non-generality, these signatures are therefore not necessary conditions of Bayesian confidence. Furthermore, they may be observed under non-Bayesian models, indicating that they are also not sufficient conditions (Fleming and Daw, 2017; Insabato et al., 2016).

A discrepancy in the literature (Navajas et al., 2017) has emerged through the confusion of divergence signature #1 with a second form, in which stimulus magnitude is replaced with measurement noise.[XII] We have shown that, while divergence signature #1 holds in some cases, there is no evidence that the second form is ever expected under the BCH, which resolves this discrepancy.

Some of our critique of the signatures has focused on the implicit assumption that experiments use non-overlapping stimulus distributions. One could object to our critique by questioning the relevance of overlapping stimulus distributions, given that non-overlapping stimulus distributions are the norm in the confidence literature (Aitchison and Latham, 2014; Kepecs and Mainen, 2012; Kepecs et al., 2008; Sanders et al., 2016). But although the work in this dissertation (Chapters 3 and 4) represents the first use of overlapping categories to study confidence, such categories have a long history in the perceptual categorization literature (Ashby and Gott, 1988; Green and Swets, 1966; Healy and Kubovy, 1981; Lee and Janke, 1964; Liu et al., 1995; Qamar et al., 2013; Sanborn et al., 2010). It has been argued that overlapping Gaussian stimulus distributions have several properties that make them more naturalistic than non-overlapping distributions (Maddox, 2002). The property most relevant here is that with overlapping categories, perfect performance is impossible, even with zero measurement noise. With overlapping categories, as in real life, identical stimuli may belong to multiple categories. Imagine a coffee drinker pouring salt rather than sugar into her drink,

[XII] Kiani et al. (2014) also note the lack of the divergence signature in their data, but because their stimuli have variable duration, optimality is more complicated to characterize (Drugowitsch et al., 2014a), and the explanation we offer here may not apply.

a child reaching for his parent's glass of whiskey instead of his glass of apple juice, or a doctor classifying a malignant tumor as benign (Augsburger et al., 2008). In all three examples, stimuli from opposing categories may be visually identical, even under zero measurement noise. For more naturalistic experiments with overlapping categories, qualitative signatures will be unusable if their derivations assume non-overlapping categories.

Given our demonstration that proposed qualitative signatures of confidence have limited applicability, what is the way forward? One option available to confidence researchers is to discover more signatures, being careful to find the specific conditions under which they are expected. Confidence experimentalists should then make sure to look for such signatures only when their tasks satisfy the specified conditions (e.g., stimulus distribution type, noise level). However, for researchers interested in testing the BCH, we do not necessarily advocate for this course of action because, even when applied to relevant experiments, the presence or absence of qualitative signatures provides an uncertain amount of evidence for or against the BCH. Testing for the presence of qualitative signatures is a weak substitute for accumulating probabilistic evidence, something that careful (Palminteri et al., 2017) quantitative model comparison does more objectively. Testing for signatures requires the experimenter to make two subjective judgments. First, the experimenter must determine whether the signature is present, a task potentially made difficult by the fact that real data is noisy. Second, the experimenter must determine how much evidence that provides in favor of the BCH, and whether further investigation is warranted. By contrast, model comparison provides a principled quantity (namely, a log likelihood) in favor of the BCH over some other model (Aitchison and Latham, 2014). Given the caveats associated with qualitative signatures, it may be that, as a field, we have no choice but to rely on formal model comparison. Chapters 3 and 4 will use model comparison to do a quantitative test of the BCH.

Chapter 3

Human confidence reports under bottom-up stimulus uncertainty

3.1 Introduction

In the previous chapter, we showed that qualitative signatures are not useful indicators of whether confidence is Bayesian, and we concluded that a formal model comparison approach may instead be required. In this and the next chapter, we will conduct a series of binary categorization tasks designed to probe the computational underpinnings of confidence.

Our tasks use simple visual stimuli in which the primary variable of interest is stimulus orientation. Many confidence studies use time-varying stimuli in which subjects are able to terminate stimulus presentation when ready to make a decision (Kiani et al., 2014; Kiani and Shadlen, 2009; van den Berg et al., 2016). However, when stimuli have variable duration, optimality is more complicated to characterize (Drugowitsch et al., 2014a); for this reason, we present stimuli for a fixed, brief amount of time. Our observer models differ by how they incorporate sensory uncertainty; therefore it is essential that we vary both the variable of interest as well as sensory uncertainty. In this chapter, we induce sensory uncertainty by manipulating external stimulus factors, specifically stimulus contrast or ellipse elongation.

After conducting our experiments, we compare the fits of Bayesian models to those of a variety of alternative models, something that is rarely done but very important for the epistemological standing of Bayesian claims (Bowers and Davis, 2012; Jones and Love, 2011). At first glance, it seems obvious that sensory uncertainty is relevant to the computation of confidence. However, this is by no means a given; in fact, a prominent proposal is that confidence is based on the distance between the measurement and the decision boundary, without any role for sensory uncertainty (Kepecs et al., 2008; Komura et al., 2013; Rahnev et al., 2011). Therefore, we test a model (Fixed) in which the response is a function of the measurement alone (equivalent to a maximum likelihood estimate of the stimulus orientation), and not of the uncertainty of that measurement (Figure 3.2, second column).

We also test heuristic models in which the subject uses their knowledge of their sensory uncertainty but does not compute a posterior distribution over category. We have previously classified such models as *probabilistic non-Bayesian* (Ma, 2012). We find that the BCH qualitatively describes human behavior but that quantitatively, even the most flexible Bayesian model is outperformed by models that take sensory uncertainty into account in a non-Bayesian way.

3.2 Methods

3.2.1 Experiment 1

During each session, each subject completed two orientation categorization tasks, Tasks A and B. On each trial, a category C was selected randomly (both categories were equally probable), and a stimulus s was drawn from the corresponding stimulus distribution and displayed. The subject categorized the stimulus and simultaneously reported their confidence on a 4-point scale, with a single button press (Figure 3.1a). Using a single button press for choice and confidence prevented post-choice influences on the confidence judgment (Navajas et al., 2016) and emphasized that confidence should reflect the observer's perception rather than a preceding motor response. The categories were defined by normal distributions on orientation, which differed by task (Figure 3.1b). In Task A, the distributions had different means ($\pm\mu_C$) and the same standard deviation (σ_C); leftward-tilting stimuli were more likely to be from category 1. Variants of Task A are common in decision-making studies (Britten et al., 1992). In Task B, the distributions had the same mean (0°) and different standard deviations (σ_1, σ_2); stimuli around the horizontal were more likely to be from category 1. Variants of Task B are less common (Liu et al., 1995; Qamar et al., 2013; Sanborn et al., 2010) but have some properties of perceptual organization tasks; for example, a subject may have to detect when a stimulus belongs to a narrow category (e.g., in which two line segments are collinear) that is embedded in a a broader category (e.g., in which two line segments are unrelated).

Subjects were highly trained on the categories; during training, we only used highest-reliability stimuli, and we provided trial-to-trial category correctness feedback. Subjects were then tested with 6 different reliability levels, which were chosen randomly on each trial. During testing, correctness feedback was withheld to avoid the possibility that confidence simply reflects a learned mapping between stimuli and the probability of being correct, something that no other confidence studies have done (Körding and Wolpert, 2004; Maloney and Mamassian, 2009; Qamar et al., 2013).

Because we are interested in subjects' intrinsic computation of confidence, we did not instruct or incentivize them to assign probability ranges to each button (e.g., by using a scoring rule (Brier, 1950; Gneiting and Raftery, 2007; Massoni et al., 2014)). If we had, we would have essentially been training subjects to use a specific model of confidence.

To ensure that our results were independent of stimulus type, we used two kinds of

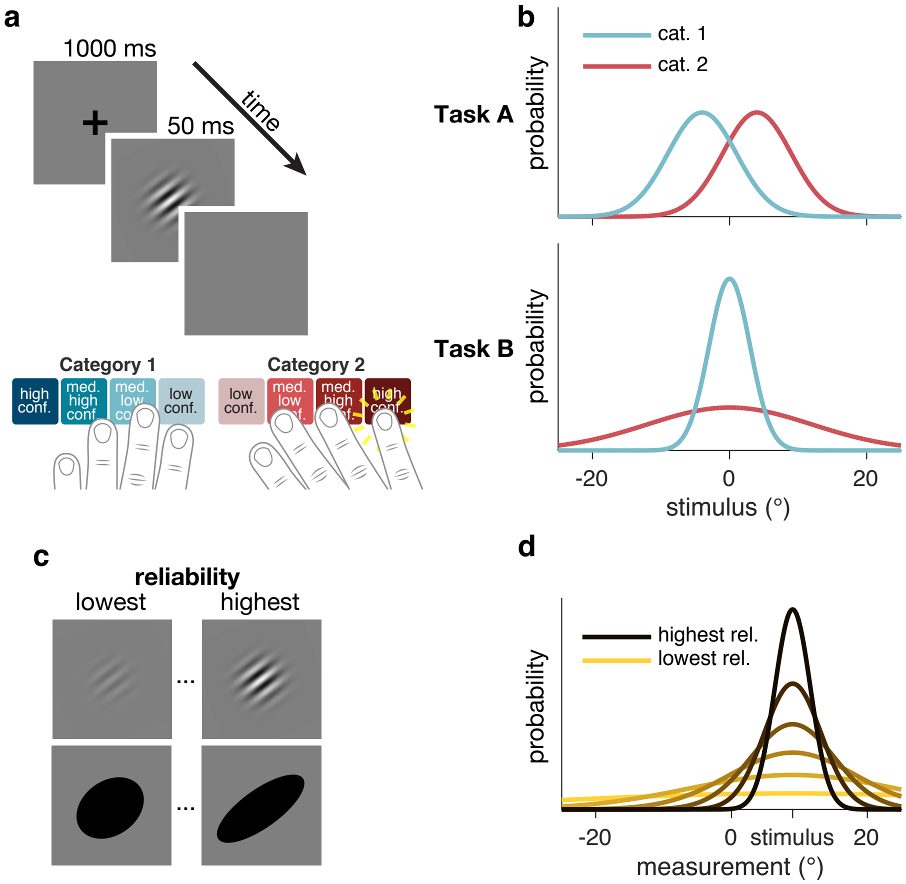

Figure 3.1 Task design. (**a**) Schematic of a test block trial. After stimulus offset, subjects reported category and confidence level with a single button press. (**b**) Stimulus distributions for Tasks A and B. (**c**) Examples of low and high reliability stimuli. Six (out of eleven) subjects saw drifting Gabors, and five subjects saw ellipses. (**d**) Example measurement distributions at different reliability levels. In all models (except Linear Neural), the measurement is assumed to be drawn from a Gaussian distribution centered on the true stimulus, with s.d. dependent on reliability.

stimuli. Some subjects saw oriented drifting Gabors; for these subjects, stimulus reliability was manipulated through contrast. Other subjects saw oriented ellipses; for these subjects, stimulus reliability was manipulated through ellipse elongation (Figure 3.1c). We found no major differences in model rankings between Gabor and ellipse subjects, therefore we will make no distinctions between the groups.

For modeling purposes, we assume that the observer's internal representation of the stimulus is a noisy measurement x, drawn from a Gaussian distribution with mean s and s.d. σ (Figure 2.1, Figure 3.1d). In the model, σ (i.e., uncertainty) is a fitted function of stimulus reliability.

3.2.1.1 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.

3.2.1.2 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 (Brainard, 1997; Pelli, 1997) 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 0.5 cycles per degrees of visual angle (dva), a speed of 6 cycles per second, a Gaussian envelope with a standard deviation of 1.2 dva, and a randomized starting phase. Each ellipse had a total area of 2.4 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 3.2.1.3).

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) (Figure 3.1b). We chose these category means and standard deviations such that the accuracy of an optimal observer would be around 80%.

3.2.1.3 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 session). 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 2880 category training trials, 240 confidence training trials, and 4320 testing 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 3.1b, 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 (Figure 3.1a): 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 (Figure 3.1b). 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 3.1a, 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 (Figure 3.1c). 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 (Sanders et al., 2016), 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 s 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.

3.2.2 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.
- Subjects completed a total of 3240 testing trials.
- 8 subjects (0 male), aged 19–23, participated. None were participants in experiment 1, and again, none were fellow scientists.
- Drifting Gabors were used; no subjects saw ellipses.

3.2.3 Experiment 3: Task B only

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

- Subjects completed blocks of Task B only.
- Subjects completed a total of 3240 testing trials.
- 15 subjects (7 female), aged 19–30, participated. None were participants in experiments 1 or 2.
- Drifting Gabors were used; no subjects saw ellipses.

3.2.4 Modeling

3.2.4.1 Measurement noise

For models (such as our core 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 variance σ^2: $\sigma^2(c) = \gamma + \alpha c^{-\beta}$. We have previously (Qamar et al., 2013) used this power law relationship because it encompasses a large family of monotonically decreasing relationships using only three parameters. The relationship is also consistent with a form of the Naka-Rushton function (DiMattina, 2016; Naka and Rushton, 1966) commonly used to describe the mapping from reliability to neural gain g: $g = \frac{\gamma c^\beta}{c^\beta + \alpha}$. The power law relationship then holds under the assumption that measurement noise variance is inversely proportional to gain (Ma et al., 2006).

For all models except the Bayesian model with additive precision, we assumed additive orientation-dependent noise in the form of a rectified 2-cycle sinusoid, accounting for the finding that measurement noise is higher at non-cardinal orientations (Girshick et al., 2011). The measurement noise s.d. comes out to

$$\sigma(c, s) = \sqrt{\gamma + \alpha c^{-\beta}} + \psi \left| \sin \frac{\pi s}{90} \right|. \tag{3.1}$$

3.2.4.2 Response probability

We coded all responses as $r \in \{1, 2, \ldots, 8\}$, 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 $p(x \mid s_i) = \mathcal{N}(x; s_i, \sigma_i^2)$ (i.e., a normal distribution over x with mean s_i and variance σ_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 distribution. We find the boundaries $(b_{r_i-1}(\sigma_i), b_{r_i}(\sigma_i))$ in measurement space, as defined by the fitting model and parameters θ, and then compute the probability mass of the measurement distribution between the boundaries:

$$p_{m,\theta}(r_i \mid s_i, \sigma_i) = \int_{b_{r_i-1}}^{b_{r_i}} \mathcal{N}(x; s_i, \sigma_i^2)\,\mathrm{d}x. \tag{3.2}$$

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:

$$\log p(\text{data} \mid \theta) = \sum_{i=1}^{t} \log p(r_i \mid \theta) = \sum_{i=1}^{t} \log p_\theta(r_i \mid s_i, \sigma_i). \tag{3.3}$$

3.2.4.3 Model specification

Bayesian A Bayes-optimal observer uses knowledge of the generative model to make a decision that maximizes the probability of being correct. Here, when the measurement on a given trial is x, this strategy amounts to choosing the category C for which the posterior probability $p(C \mid x)$ is highest. This is equivalent to reporting category 1 when the log posterior ratio, $d = \log \frac{p(C=1|x)}{p(C=2|x)}$, is positive.

In Task A, d is $d_\text{A} = \frac{2x\mu_C}{\sigma^2+\sigma_C^2}$. Therefore, the ideal observer reports category 1 when x is positive; this is the structure of many psychophysical tasks (Green and Swets, 1966). In Task B, however, d is $d_\text{B} = \frac{1}{2}\log\frac{\sigma^2+\sigma_2^2}{\sigma^2+\sigma_1^2} - \frac{\sigma_2^2-\sigma_1^2}{2(\sigma^2+\sigma_1^2)(\sigma^2+\sigma_2^2)}x^2$; the observer needs both x and σ in order to make an optimal decision.

From the point of view of the observer, σ is the trial-to-trial level of sensory uncertainty associated with the measurement (Ma, 2010). In a minor variation of the optimal observer, we allow for the possibility that the observer's prior belief over category, $p(C)$, is different from the true value of $(0.5, 0.5)$; this adds a constant to d_A and d_B.

We introduce the Bayesian confidence hypothesis (BCH), stating that confidence reports depend on the internal representation of the stimulus (here x) only via d. In the BCH, the observer chooses a response by comparing d to a set of category and confidence boundaries. For example, whenever d falls within a certain range, the observer presses the "medium-low confidence, category 2" button. The BCH is thus an extension of the choice model described above, wherein the value of d is used to compute confidence as well as chosen category. There is another way of thinking about this. Bayesian models assume that subjects compute d in order to make an optimal choice. Assuming people compute d at all, are they able to use it to report confidence as well? We refer to the Bayesian model here as simply "Bayes." We also tested several more constrained versions of this model.

The observer's decision can be summarized as a mapping from a combination of a measurement and an uncertainty level (x, σ) to a response that indicates both category and confidence. We can visualize this mapping as in Figure 3.2, first column. It is clear that the pattern of decision boundaries in the BCH is qualitatively very different between Task A and Task B. In Task A, the decision boundaries are quadratic functions of uncertainty; confidence decreases monotonically with uncertainty and increases with the distance of the measurement from 0. In Task B, the decision boundaries are neither linear nor quadratic.

Derivation of d_A *and* d_B. The log posterior ratio d is equivalent to the log likelihood ratio plus the log prior ratio:

$$d = \log \frac{p(C = 1 \mid x)}{p(C = 2 \mid x)} = \log \frac{p(x \mid C = 1)}{p(x \mid C = 2)} + \log \frac{p(C = 1)}{p(C = 2)}. \tag{3.4}$$

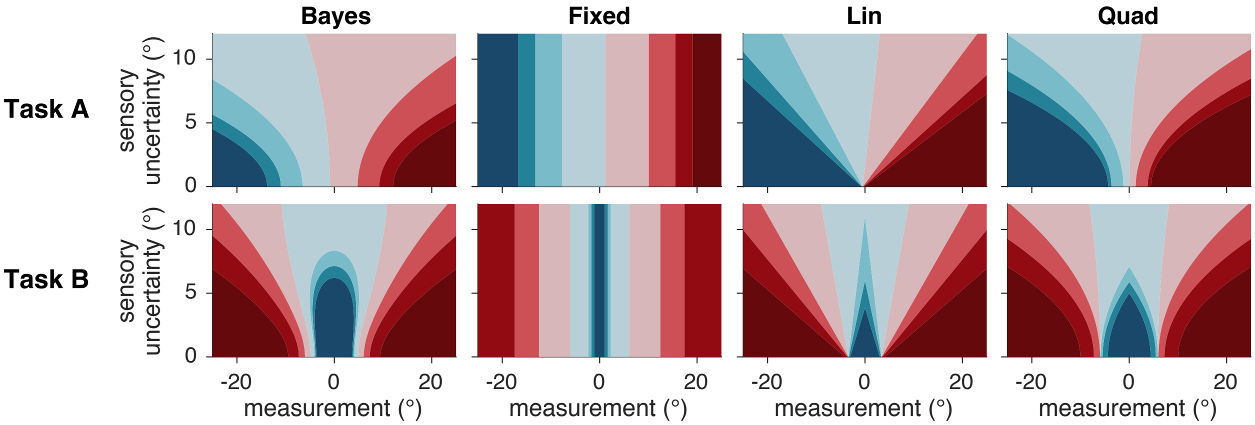

Figure 3.2 Decision rules/mappings in four models. Each model corresponds to a different mapping from a measurement and uncertainty level to a category and confidence response. Colors correspond to category and confidence response, as in Figure 3.1a. Plots were generated from the mean of subject 4's posterior distribution over parameters. These figures should be used only as an aid for understanding the models' decision rules, not for closely interpreting the different fitted rules across models; interpretation is complicated by, among other considerations, the fact that some regions have very few trials.

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:

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

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

$$p(x \mid C) = \int \mathcal{N}(s; x, \sigma^2) \mathcal{N}(s; \mu_C, \sigma_C^2) \, \mathrm{d}s = \mathcal{N}(x; \mu_C, \sigma^2 + \sigma_C^2). \tag{3.5}$$

Plugging the task- and category-specific μ_C and σ_C into Equation (3.5), and substituting the resulting expression back into Equation (3.4), we get:

$$d_{\mathrm{A}} = \frac{2x\mu_1}{\sigma^2 + \sigma_1^2} + \log \frac{p(C=1)}{p(C=2)} \tag{3.6}$$

$$d_{\mathrm{B}} = \frac{1}{2} \log \frac{\sigma^2 + \sigma_2^2}{\sigma^2 + \sigma_1^2} - \frac{\sigma_2^2 - \sigma_1^2}{2(\sigma^2 + \sigma_1^2)(\sigma^2 + \sigma_2^2)} x^2 + \log \frac{p(C=1)}{p(C=2)}. \tag{3.7}$$

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.

Because the decision boundaries are free parameters, our models effectively include a large family of possible cost functions. A different cost function would be equivalent to a rescaling of the confidence boundaries $\mathbf{k}$. To see this, it is probably easiest to consider category choice alone; there, asymmetric costs for getting either category wrong would translate into a different value of k_4, the category decision boundary (i.e., the observer's prior over category). For us, this boundary (like all other boundaries) is a free parameter.

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

Levels of strength. We formulated several levels of strength of the Bayesian model, with weaker versions having fewer assumptions and more sets of mappings between the posterior probability of being correct and the confidence report (Figure 3.3).

In $\text{Bayes}_{\text{Ultrastrong}}$, $\mathbf{k}$ is symmetric across k_4: $k_{4+j} - k_4 = k_4 - k_{4-j}$ for $j \in \{1, 2, 3\}$. Furthermore, in $\text{Bayes}_{\text{Ultrastrong}}$, $\mathbf{k}_{\mathrm{A}} = \mathbf{k}_{\mathrm{B}}$. So $\text{Bayes}_{\text{Ultrastrong}}$ has a total of 4 free boundary parameters: k_1, k_2, k_3, k_4. $\text{Bayes}_{\text{Ultrastrong}}$ consists of the observer determining the response by comparing d_{A} and d_{B} to a single symmetric set of boundaries (Figure 3.3, left column).

$\text{Bayes}_{\text{Strong}}$ is identical to $\text{Bayes}_{\text{Ultrastrong}}$ except that $\mathbf{k_A}$ is allowed to differ from $\mathbf{k_B}$. So $\text{Bayes}_{\text{Strong}}$ has a total of 8 free boundary parameters: $k_{1\text{A}}, k_{2\text{A}}, k_{3\text{A}}, k_{4\text{A}}, k_{1\text{B}}, k_{2\text{B}}, k_{3\text{B}}, k_{4\text{B}}$. $\text{Bayes}_{\text{Strong}}$ consists of the observer determining the response by comparing d_{A} to a symmetric set of boundaries, and d_{B} to a different symmetric set of boundaries (Figure 3.3, middle 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: $k_{1\text{A}}, k_{2\text{A}}, k_{3\text{A}}, k_{4\text{A}}, k_{1\text{B}}, k_{2\text{B}}, k_{3\text{B}}, k_{4\text{B}}, k_{5\text{B}}, k_{6\text{B}}, k_{7\text{B}}$. $\text{Bayes}_{\text{Weak}}$ consists of the observer comparing d_{A} to a symmetric set of boundaries, and d_{B} to a different non-symmetric set of boundaries (Figure 3.3, right column).

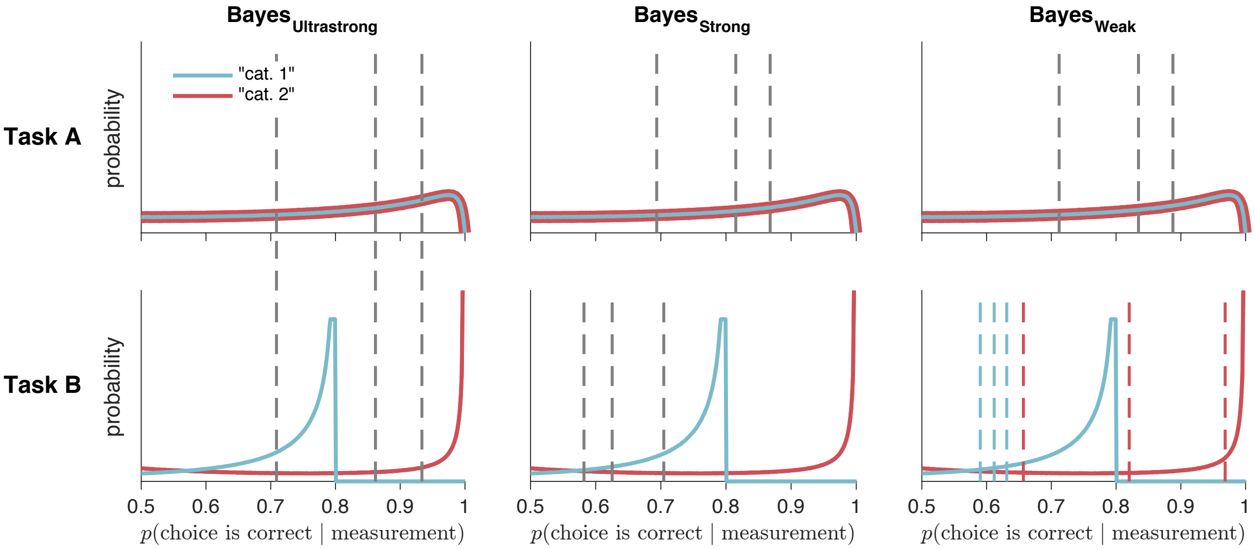

Figure 3.3 Distributions of posterior probabilities of being correct, with confidence criteria for Bayesian models with three different levels of strength. Solid lines represent the distributions of posterior probabilities for each category and task in the absence of measurement noise. Dashed lines represent 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 $\text{Bayes}_{\text{Weak}}$, as in the non-Bayesian models, 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 3.2.

Decision boundaries. In the Bayesian models without d noise, we translate boundary parameters $\mathbf{k}$ to measurement boundaries $\mathbf{b}$ corresponding to fitted noise levels σ. To do this, we use the parameters $\mathbf{k}$ as the left-hand side of Equations (3.6) and (3.7) 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 (3.6) and (3.7) 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 (Acerbi et al., 2012). We then computed 101 response probabilities for each trial (Section 3.2.4.2), one for each draw of η_d, and computed the weighted average according to $p(\eta_d)$.

Probability correct with additive precision We tested a model in which the decision variable was a weighted mixture of precision (equivalent in this case to the Fisher information of the measurement variable x) and the perceived probability of being correct (Navajas et al., 2017). In this model, the decision variable is $\frac{\omega}{\sigma^2} + \frac{1}{1+\exp(-|d|)}$, where ω is a free parameter. To find the measurement boundaries $\mathbf{b}(\sigma)$, we substituted Equations (3.6) and (3.7) for d, and set the whole value equal to parameters $\mathbf{k}$, solving for x at the fitted levels of σ. This model can be considered a hybrid Bayesian-heuristic model. Like $\text{Bayes}_{\text{Ultrastrong}}$, it has 4 free boundary parameters. Although the model is a hybrid Bayesian-heuristic model, not a strictly Bayesian one, we refer to it as $\text{Bayes}_{\text{Ultrastrong}}$ + precision in Figure 3.13.

Fixed In Fixed, the observer compares the measurement to a set of boundaries that are not dependent on σ (Figure 3.2, second column). We fit free parameters $\mathbf{k}$, and use measurement boundaries $b_r = k_r$.

Lin and Quad We derived two additional probabilistic non-Bayesian models, Lin and Quad, from the observation that the Bayesian decision criteria are an approximately linear function of uncertainty in some measurement regimes and approximately quadratic in others. These models are able to produce approximately Bayesian behavior without actually performing any computation of the posterior. In Lin and Quad, subjects base their response on a linear or a quadratic function of x and σ, respectively. A comparison of the Lin and Quad columns to the Bayes column in Figure 3.2 demonstrates that Lin and Quad can approximate the Bayesian mapping from (x, σ) to response despite not being based on the Bayesian decision variable.

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).

Lin and Quad are each a supermodel of Fixed. In other words, there are parameter settings where Lin and Quad are equivalent to Fixed (although our model comparison methods ensure that the models are still distinguishable, see Section 3.3.5). Additionally, in Task A, Quad is a supermodel of the Bayesian models without d noise.

Orientation Estimation In Orientation Estimation, the observer uses the mixture of the two stimulus distributions as a prior distribution to compute a maximum a posteriori estimate of the stimulus:

$$\begin{aligned}\hat{s} &= \underset{s}{\operatorname{argmax}}\, p(s \mid x) \\ &= \underset{s}{\operatorname{argmax}}\, p(x \mid s)p(s) \\ &= \underset{s}{\operatorname{argmax}} \left[\mathcal{N}(s; x, \sigma^2)(p(s \mid C = 1) + p(s \mid C = 2))\right]. \end{aligned} \tag{3.8}$$

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 (github.com/lacerbi/gmm1) 1-D Gaussian mixture model toolbox to solve Equation (3.8), computing a lookup table containing the value of $\hat{s}$ as a function of x and σ (Acerbi et al., 2014). 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 $\mathbf{b}(\sigma)$.

Linear Neural In Linear Neural, subjects base their response on a linear function of the output of a hypothetical population of neurons.

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 $ag \int \tilde{s} \exp\left(-\frac{(s-\tilde{s}_j)^2}{2\sigma^2_{\text{TC}}}\right) \mathrm{d}\tilde{s} = ags\sqrt{2\pi\sigma^2_{\text{TC}}}$. The variance of z is $\sum_j w_j^2 f_j(s) = ag \int \tilde{s}^2 \exp\left(-\frac{(s-\tilde{s})^2}{2\sigma^2_{\text{TC}}}\right) \mathrm{d}\tilde{s} = ag\sqrt{2\pi\sigma^2_{\text{TC}}}(\sigma^2_{\text{TC}} + s^2)$.

Now that we have the mean and variance of z, we 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 (3.2) with

$$p_\theta(r_i \mid s_i, \sigma_i) = \int_{k_{r_i-1}}^{k_{r_i}} \mathcal{N}(z; ags_i\sqrt{2\pi\sigma_{\mathrm{TC}}^2}, ag\sqrt{2\pi\sigma_{\mathrm{TC}}^2}(\sigma_{\mathrm{TC}}^2 + s_i^2))\,\mathrm{d}z.$$

3.2.4.4 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 as the above "repeat lapse" λ_{repeat}.

3.2.4.5 Parameterization

Because of tradeoffs when directly fitting parameters γ, α, β, we re-parameterized Equation (3.1) as

$$\sigma(c, s) = \sqrt{\sigma_{\mathrm{L}}^2 + \frac{(\sigma_{\mathrm{L}}^2 - \sigma_{\mathrm{H}}^2)(c^{-\beta} - c_{\mathrm{L}}^{-\beta})}{c_{\mathrm{L}}^{-\beta} - c_{\mathrm{H}}^{-\beta}} + \psi\left|\sin\frac{\pi s}{90}\right|},$$

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, and β was a free parameter 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 (3.1) was replaced with free s.d. parameters ($\sigma_{\text{rel. }1}, \dots, \sigma_{\text{rel. }6}$) corresponding to each of the six reliability levels.

For models where subjects had incorrect knowledge about their measurement noise, we fit two sets of uncertainty-related parameters. One set was for the generative noise (used in Equation (3.2)), and the other set was for the subject's believed noise (used in Equations (3.6) to (3.8)).

All parameters that defined the width of a distribution (e.g., $\sigma_{\mathrm{L}}, \sigma_{\mathrm{H}}, \sigma_d, \sigma_{\text{rel. }1}, \dots$) were sampled in log-space and exponentiated during the computation of the log likelihood.

3.2.4.6 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 both for model comparison and for plotting model fits. We used the log posterior

$$\log p(\theta \mid \text{data}) = \log p(\text{data} \mid \theta) + \log p(\theta) + \text{constant}, \tag{3.9}$$

where $\log p(\text{data} \mid \theta)$ is given in Equation (3.3). We assumed a factorized prior over each parameter j:

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

where j is the parameter index and n is the number of parameters. We took uniform (or, for parameters that were standard deviations, log-uniform) priors over reasonable, sufficiently large ranges (Acerbi et al., 2014), which we chose before fitting any models.

We sampled from the probability distribution using a Markov Chain Monte Carlo (MCMC) method, slice sampling (Neal, 2003). 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) from the posterior distribution over parameters. We discarded the first third of the samples and kept 6,667 of the remaining samples, evenly spaced to reduce autocorrelation. All samples with log posteriors more than 40 below the maximum log posterior were discarded. Marginal probability distributions of the sample log likelihoods were visually checked for convergence across chains. In total we had 842 model and dataset combinations, with a median of 26,668 kept samples (IQR = 13,334).

After sampling, we conducted a visual check to confirm that our parameter ranges were sufficiently large. For each model, we plotted the posterior distribution over parameter values for each subject; an example plot is shown in Figure 3.4. Visual checks of these plots confirmed that the distributions are unimodal and roughly Gaussian. Visual checks also confirmed that the parameter distributions are well-contained within the chosen parameter ranges, except for the distributions of:

- Lapse rate parameters, which tend to mass around 0, where they are necessarily bounded.
- Log noise parameters, which have a large negative range where they are effectively at zero noise.
- Upper confidence boundary parameters, which become small for subjects who frequently report "high confidence," or large for subjects who frequently do.

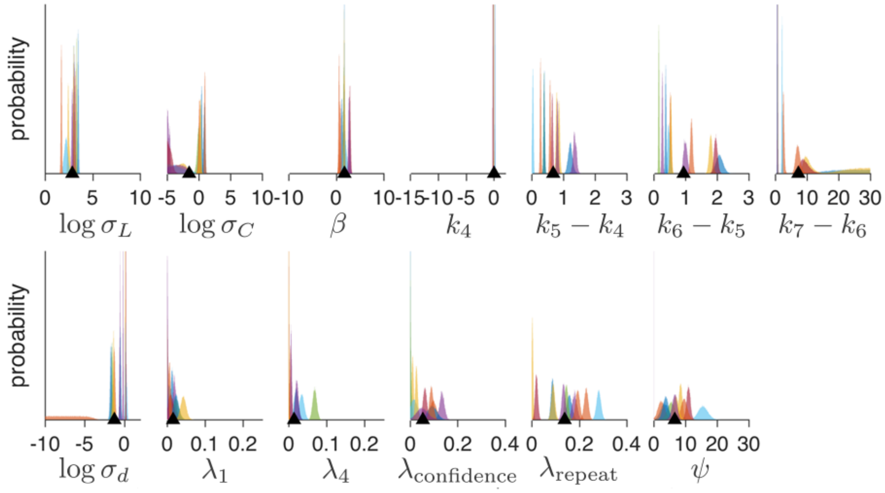

Figure 3.4 Posterior distributions over parameter values for an example model. Each subplot represents a parameter of the model. Each colored histogram represents the sampled posterior distribution for a parameter and a subject in experiment 1, with colors consistent for each subject. The limits of the x-axis indicates the allowable range for each parameter. Black triangles indicate the overall mean parameter value.

3.2.4.7 Model comparison

Model groupings We used 8 groupings of model-subject combinations where it made sense to consider the models as being on equal footing for the purpose of model comparison. The model-subject combinations were grouped by: experiment (which corresponded to subject population), data type (category response only vs. category and confidence response), task type (Task A, B, or both fit jointly). The 8 groupings correspond to Figures 3.13 to 3.20 and Tables 3.1 to 3.7.

Metric choice 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.

Most information criteria (such as AIC, BIC, and AICc) are based on a point estimate for θ, typically θ_{MLE}, the θ that maximizes the log likelihood of the dataset (Equation (3.3)). For instance, AIC adds a correction for the number of parameters n to the log likelihood of the dataset: $\text{AIC} = -2\sum_{i=1}^{t} \log p(r_i \mid \theta_{\text{MLE}}) + 2n$.

WAIC is a more Bayesian approach to information criteria that adds a correction for the effective number of parameters (Gelman et al., 2013). Because WAIC is based on samples from the full posterior of θ (Equation (3.9), typically sampled via MCMC), it takes into account the model's uncertainty landscape.

Although information criteria are computationally convenient, they are based on asymptotic results and assumptions about the data that may not always hold (Gelman et al., 2013). 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 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 (PSIS-LOO, referred to here simply as LOO) proposed by Vehtari et al. (2015) for approximating leave-one-out cross-validation that, like WAIC, uses samples from the full posterior of θ:

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

where θ_u is the u-th sampled set of parameters, and $w_{i,u}$ is the importance weight of trial i for sample u. Pareto smoothed importance sampling provides an accurate and reliable estimate of the weights. LOO is currently the most accurate approximation of leave-one-out cross-validation (Acerbi et al., 2017). Conveniently, it has a natural diagnostic that allows the user to know when the metric may be inaccurate (Vehtari et al., 2015); we used that diagnostic and confirmed that our use of the metric is justified.

Metric aggregation *Summed LOO differences.* In all figures where we present model comparison results (e.g., Figure 3.10, right column), we aggregate LOO scores by the following procedure. Choose a reference model (usually the one with the lowest mean LOO score across subjects). Subtract all LOO scores from the corresponding subject's score for the reference model; this converts all scores to a LOO "difference from reference" score, with higher scores being worse. Repeat the following standard bootstrap procedure 10,000 times: Choose randomly, with replacement, a group of datasets equal to the total number of unique datasets, and take the sum over subjects of their "difference from reference" scores for each model. Plots indicate the median and 95% CI of these bootstrapped summed "difference from reference" scores. This approach implicitly assumes that all data was generated from the same model.

To confirm that our sample size was large enough to trust our bootstrapped confidence intervals, we bootstrapped our bootstrapping procedure to see how the confidence intervals were affected by the number of subjects N. For an example pair of models that we might be interested in comparing, and took the 11 LOO differences between the models, one for each subject in experiment 1. For each N between 2 and 11, we took 50 subsamples of our subject LOO differences with replacement; this is akin to running the experiment 50 times for each N. For each subsample, we conducted the above bootstrap procedure, which give us a median and 95% CI on the mean of differences. We then plot the mean of these values, with error bars indicating ±1 s.d., at each N (Figure 3.5a). A visual check indicates that the confidence interval appears to converge at about $N = 9$. This indicates that our bootstrapped confidence intervals are trustworthy.

Group level Bayesian model selection. We also used LOO scores to compute two metrics that allow for model heterogeneity across the group. The first metric was "protected exceedance probability," the posterior probability that one model occurs more frequently than

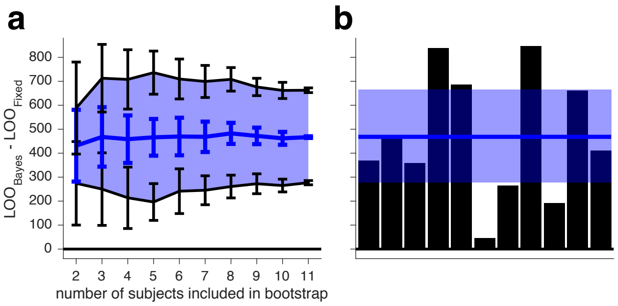

Figure 3.5 Example analysis of a bootstrapped confidence interval. (**a**) Uncertainty estimates for bootstrapped confidence intervals, as a function of the number of subjects included. Blue line represents the median bootstrapped mean of LOO differences, and black lines indicate the lower and upper bounds of the 95% CI. Error bars represent ±1 s.d. (**b**) For comparison to **a**, the standard style of plot used to show model comparison results (e.g., Figure 3.9).

any other model in the set (Rigoux et al., 2014), above and beyond chance (e.g., Figure 3.13b). The second was the expected posterior probability that a model generated the data of a randomly chosen dataset (Stephan et al., 2009) (e.g., Figure 3.13c). The latter metric assumes a uniform prior over models, which is a function of the total number of datasets. We used the SPM12 (www.fil.ion.ucl.ac.uk/spm) software package to compute these metrics.

In all but one of the 8 model groupings, all three methods of metric aggregation identify the same overall best model. For example, in Figure 3.13, one model (Quad + non-param. σ) has the lowest summed LOO differences, the highest protected exceedance probability, and the highest expected posterior probability.

3.2.4.8 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 mean ±1 s.e.m. of synthetic group datasets.

For plots with orientation on the horizontal axis (e.g., Figure 3.8j–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.

3.3 Results

Since our models do not include any learning effects, we wanted to ensure that task performance was stable. For all tasks and experiments, we found no evidence that performance changed significantly as a function of the number of trials. For each experiment and task (the 5 lines in Figure 3.6), we fit a logistic regression to the binary correctness data for each subject, obtaining a set of slope coefficients. We then used a t-test to determine whether these sets of coefficients differed significantly from zero. In no group did the slopes differ significantly from zero; across all 5 groupings the minimum p-value was 0.077 (Task A, experiment 2), which would not be significant even before correcting for multiple comparisons.

3.3.1 Descriptive statistics (experiment 1)

Each trial consists of the experimentally determined orientation and reliability level and the subject's category and confidence response (an integer between 1 and 8). This is a very rich data set. Briefly, we find the following effects: performance and confidence increase as a function of reliability (Figure 3.8a,b,h,i), and high-confidence reports are less frequent than

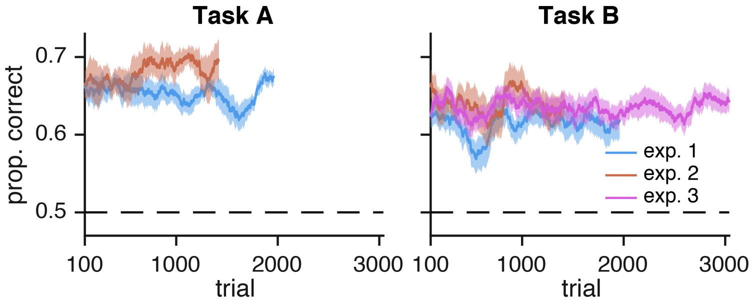

Figure 3.6 Performance as a function of number of trials, for both tasks and for all experiments. Performance was computed as a moving average over test trials (200 trials wide). Shaded regions represent ±1 s.e.m. over subjects. Performance did not change significantly over the course of each experiment.

low-confidence reports (Figure 3.8e,f). Note Figure 3.8c,d especially; this is the projection of the data that we will use to demonstrate model fits for the rest of this chapter. We use this projection because the vertical axis (mean button press) most closely approximates the form of the raw data. Additionally, because our models are differentiated by how they use uncertainty, it is informative to plot how response changes as a function of reliability, in addition to category and task.

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 ($F_{5,50} = 59.6, p < 10^{-15}$) (Figure 3.8a).

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 ($F_{5,50} = 28, p < 10^{-12}$) (Figure 3.8b).

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 ($F_{5,50} = 58.7, p < 10^{-13}$) (Figure 3.8c).

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 ($F_{5,50} = 29.2, p < 10^{-8}$) (Figure 3.8d).

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 ($F_{3,30} = 1.45, p = 0.25$) (Figure 3.8e).

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 (Figure 3.8f).

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 ($F_{3,30} = 0.878, p = 0.436$) (Figure 3.8g).

Figure 3.8l,m shows psychometric choice curves for both tasks, at all 6 levels of reliability. Each point represents roughly the same number of trials.

Figure 3.8n,o shows a similar set of psychometric curves. These curves differ from Figure 3.8l,m in that they represent the mean button press rather than mean category choice.

In Task A (Figure 3.8l,n), mean category choice and mean button press depend monotonically on orientation, with a slope that increases with reliability. In Task B (Figure 3.8m,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.

Reaction times did not vary with stimulus characteristics or with response, suggesting that drift-diffusion models would not provide more explanatory power for our dataset than the static models that we use.

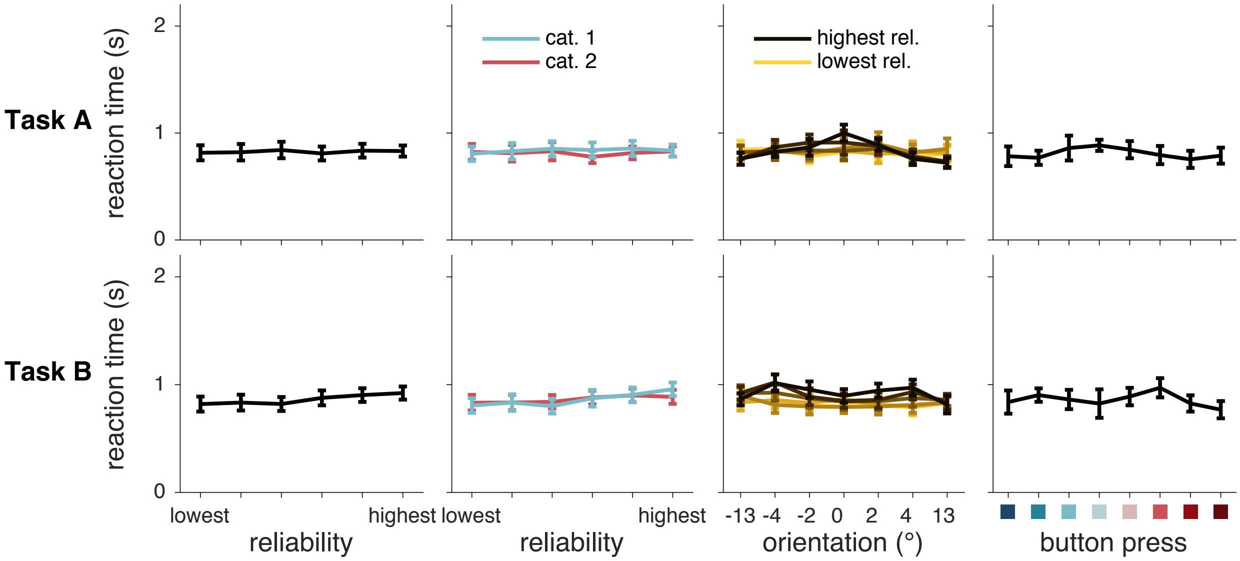

Figure 3.7 Reaction times are constant as a function of: reliability (first column); reliability and true category (second column); orientation and reliability (second column); button press (fourth column).

3.3.2 Model comparison

We used Markov Chain Monte Carlo (MCMC) sampling to fit models to raw individual-subject data. To account for overfitting, we compared models using leave-one-out cross-validated log likelihood scores (LOO) computed with the full posteriors obtained through MCMC (Vehtari et al., 2015). A model recovery analysis ensured that our models are meaningfully

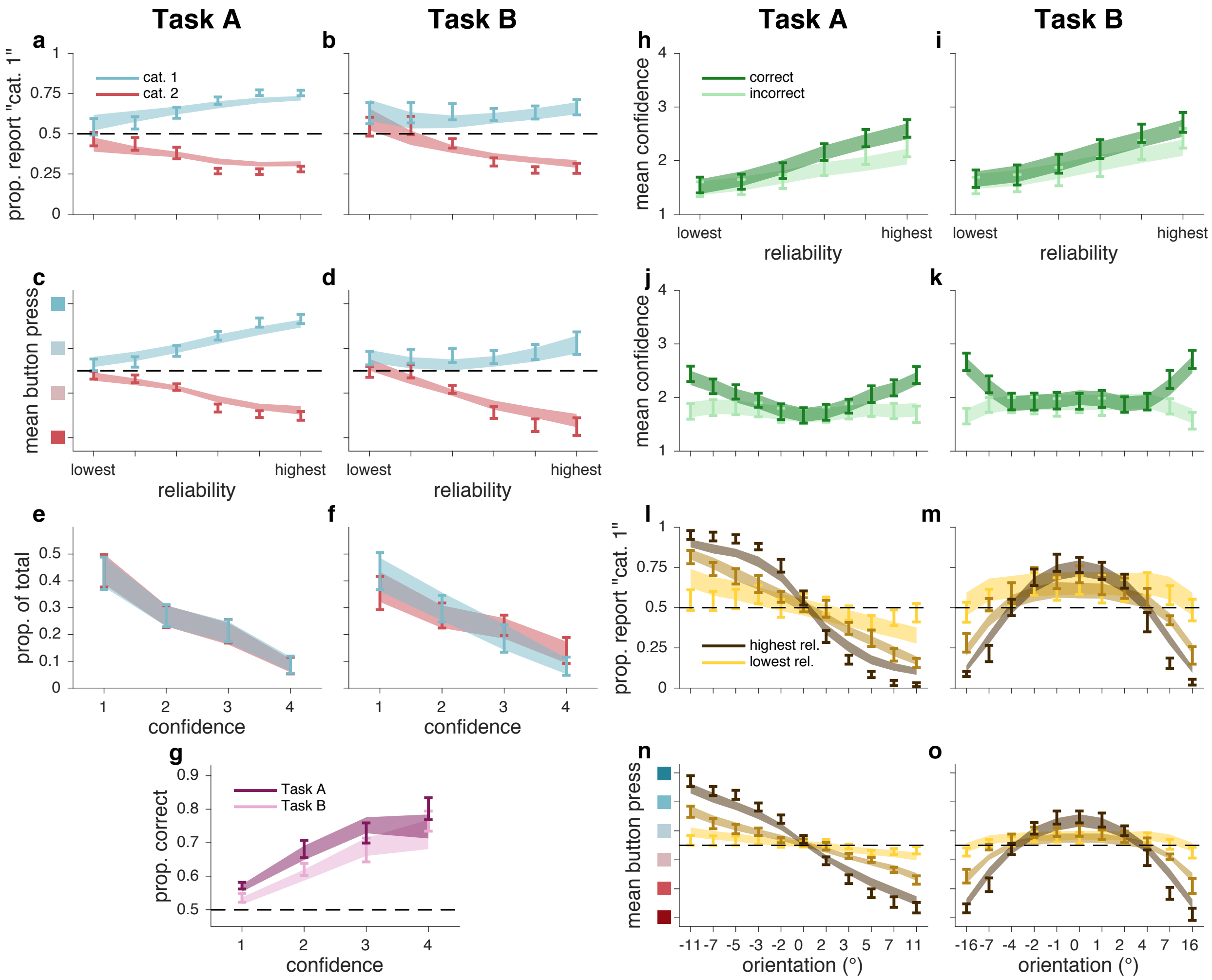

Figure 3.8 Behavioral data and fits from best model (Quad), experiment 1. Error bars represent ±1 s.e.m. across 11 subjects. Shaded regions represent ±1 s.e.m. on model fits (). (**a,b**) Proportion "category 1" reports as a function of stimulus reliability and true category. (**c,d**) Mean button press as a function of stimulus reliability and true category. (**e,f**) Normalized histogram of confidence reports for both true categories. (**g**) Proportion correct category reports as a function of confidence report and task. (**h,i**) Mean confidence as a function of stimulus reliability and correctness. (**j,k**) Mean confidence as a function of stimulus orientation and reliability. (**l,m**) Proportion "category 1" reports as a function of stimulus orientation and reliability. (**n,o**) Mean button press as a function of stimulus orientation and reliability. (**c,d,n,o**) Vertical axis label colors correspond to button presses, as in Figure 3.1a. (**l–o**) For clarity, only 3 of 6 reliability levels are shown, although models were fit to all reliability levels.

distinguishable (Figure 3.26). Unless otherwise noted, models were fit jointly to Task A and B category and confidence responses.

Use of sensory uncertainty. We first compared Bayes to the Fixed model, in which the observer does not take trial-to-trial sensory uncertainty into account (Figure 3.9). Fixed provides a poor fit to the data, indicating that observers use not only a point estimate of their measurement, but also their uncertainty about that measurement. Bayes outperforms Fixed by a summed LOO difference (median and 95% CI of bootstrapped sums across subjects) of 2265 [498, 4253]. For the rest of this chapter, we will report model comparison results using this format (Section 3.2.4.7).

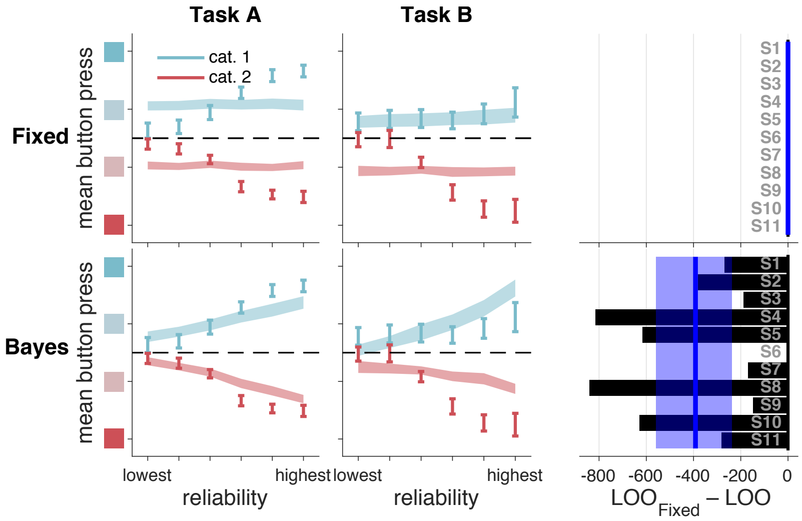

Figure 3.9 Model fits and model comparison for models Fixed and Bayes. Bayes provides a better fit, but both models have large deviations from the data. Left and middle columns: model fits to mean button press as a function of reliability, true category, and task. Error bars represent ±1 s.e.m. across 11 subjects. Shaded regions represent ±1 s.e.m. on model fits, with each model on a separate row. Right column: LOO model comparison. Bars represent individual subject LOO scores for Bayes, relative to Fixed. Negative (leftward) values indicate that, for that subject, Bayes had a higher (better) LOO score than Fixed. Blue lines and shaded regions represent, respectively, medians and 95% CI of bootstrapped mean LOO differences across subjects. These values are equal to the summed LOO differences reported in the text divided by the number of subjects.

Although Bayes fits better than Fixed, it still shows systematic deviations from the data, especially at high reliabilities. (Because we fit our models to all of the raw data and because boundary parameters are shared across all reliability levels, the fit to high-reliability trials is constrained by the fit to low-reliability trials.)

Noisy log posterior ratio. To see if we could improve Bayes's fit, we tried a version that included decision noise, i.e. noise on the log posterior ratio d. We assumed that this noise takes the form of additive zero-mean Gaussian noise with s.d. σ_d. This is almost equivalent to the probability of a response being a logistic (softmax) function of d (Keshvari et al., 2012). Adding d noise improves the Bayesian model fit by 804 [510, 1134].

For the rest of the reported fits to behavior, we will only consider this version of Bayes with d noise, and will refer to this model as Bayes-dN. We will refer to Bayes-dN, Fixed, Orientation Estimation, Linear Neural, Lin, and Quad, when fitted jointly to category and confidence data from Tasks A and B, as our core models.

Heuristic models. Orientation Estimation performs worse than Bayes-dN by 2041 [385, 3623] (Figure 3.10, second row). The intuition for one way that this model fails is as follows: at low levels of reliability, the MAP estimate is heavily influenced by the prior and tends to be very close to the prior mean (0°). This explains why, in Task B, there is a bias towards reporting "high confidence, category 1" at low reliability. Linear Neural performs about as well as Bayes-dN, with summed LOO differences of 1188 [-588, 2704], and the fits to the summary statistics are qualitatively poor (Figure 3.10, third row).

Finally, Lin and Quad outperform Bayes-dN by 1398 [571, 2644] and 1667 [858, 2698], respectively. Both models provide qualitatively better fits, especially at high reliabilities (compare Figure 3.10, first row, to Figure 3.10, fourth and fifth rows), and strongly tilted orientations (compare Figure 3.21n,o to Figure 3.25n,o and Figure 3.8n,o).

We summarize the performance of our core models in Figure 3.11. Noting that a LOO difference of more than 5 is considered to be very strong evidence (Kass and Raftery, 1995), the heuristic models Lin and Quad perform much better than Bayes-dN. Furthermore, we can decisively rule out Fixed. We will now describe variants of our core models.

Non-parametric relationship between reliability and σ. One potential criticism

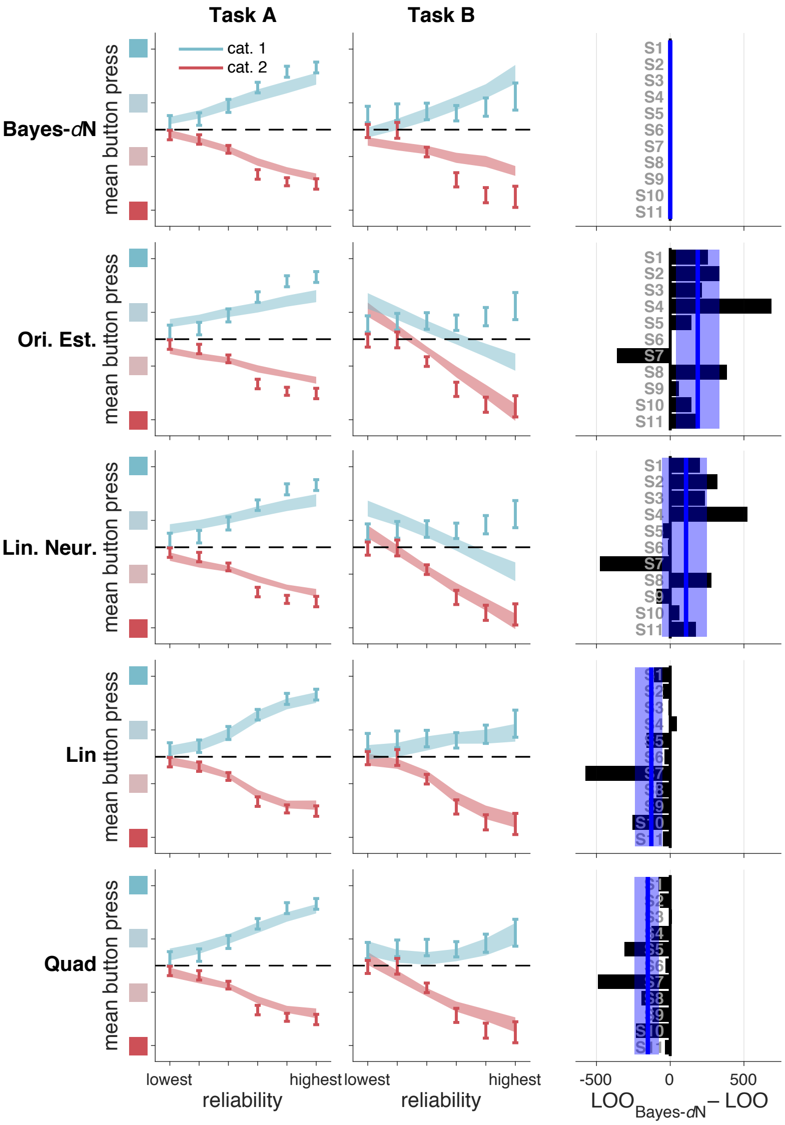

Figure 3.10 Model fits and model comparison for Bayes-dN and heuristic models. In both tasks, Bayes-dN fails to describe the data at high reliabilities; Lin and Quad provides a good fit at most reliabilities. Left and middle columns: as in Figure 3.9. Right column: bars represent individual subject LOO scores for each model, relative to Bayes-dN. Negative (leftward) values indicate that, for that subject, the model in the corresponding row had a higher (better) LOO score than Bayes-dN. Blue lines and shaded regions: as in Figure 3.9.

of our fitting procedure is that we assumed a parameterized relationship between reliability and σ. To see if our results were dependent on that assumption, we modified the models such that σ was non-parametric (i.e., there was a free parameter for σ at each level of reliability).

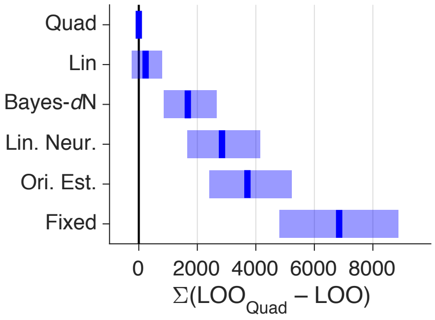

Figure 3.11 Comparison of core models, experiment 1. Models were fit jointly to Task A and B category and confidence responses. Blue lines and shaded regions represent, respectively, medians and 95% CI of bootstrapped summed LOO differences across subjects. LOO differences for these and other models are shown in Figure 3.13a.

With this feature added to our core models, Quad still fits better than Bayes-dN by 1676 [839, 2730] and it fits better than Fixed by 6097 [4323, 7901]. This feature improved Quad's performance by 325 [141, 535]. For the rest of this paper, we will only report the fits of Bayes-dN, the best-fitting non-Bayesian model, and Fixed. See supplementary figures and tables for all other model fits.

Incorrect assumptions about the generative model. Suboptimal behavior can be produced by optimal inference using incorrect generative models, a phenomenon known as "model mismatch" (Acerbi et al., 2014; Beck et al., 2012; Orhan and Jacobs, 2014). Up to now, Bayes-dN has assumed that observers have accurate knowledge of the parameters of the generative model. To test whether this assumption prevents Bayes-dN from fitting the data well, we tested a series of Bayesian models in which the observer has inaccurate knowledge of the generative model.

Bayes-dN assumed that, because subjects were well trained, they knew the true values of σ_C, σ_1, and σ_2, the standard deviations of the stimulus distributions. We tested a model in which these values were free parameters, rather than fixed to the true value. We would expect these free parameters to improve the fit of Bayes-dN in the case where subjects were not trained enough to sufficiently learn the stimulus distributions. This feature improves Bayes-dN's fit by 908 [318, 1661], but it still underperforms Quad by 768 [399, 1144].

Previous models also assumed that subjects had full knowledge of their own measurement noise; the σ used in the computation of d was identical to the σ that determined their measurement noise. We tested models in which we fit $\sigma_{\text{measurement}}$ and $\sigma_{\text{inference}}$ as two independent functions of reliability (Acerbi et al., 2014). This feature improves Bayes-dN's fit by 1310 [580, 2175], but it still underperforms Quad by 362 [162, 602].

Levels of strength of the Bayesian model. The Bayesian model is unique in that it is possible to formulate a principled version with relatively few boundary parameters. In principle, it is possible that such a model could perform better than weaker, more flexible models, if those models are overfitting. The previously described Bayesian model, which we will temporarily refer to as Bayes$_{\text{Weak}}$, has many free boundary parameters, making relatively few assumptions about the mappings between the posterior probability of being correct and the confidence report (Figure 3.3). We formulated two stronger versions of the BCH. The *strong BCH* assumes that boundary parameters are fixed across both categories. The *ultrastrong BCH* additionally assumes that boundary parameters are fixed across tasks A and B.

Most studies cannot distinguish between the strong and ultrastrong BCH because they test subjects in only one task. Furthermore, the weak BCH is only justifiable in tasks where the categories have different distributions of the posterior probability of being correct; the subject may then rescale their mappings between the posterior and their confidence. Here, one can see that Task B has this feature by observing that, in the bottom row of Figure 3.3, the distributions of posterior probabilities are different for the two categories). Most experimental tasks are like Task A, where the distributions are identical. We compared our previously described, weak Bayesian model, to Bayes$_{\text{Ultrastrong}}$-dN, Bayes$_{\text{Strong}}$-dN, models that make these stronger assumptions.

Bayes$_{\text{Strong}}$-dN and Bayes$_{\text{Ultrastrong}}$-dN each underperform Bayes$_{\text{Weak}}$-dN by 819 [441, 1369] and 2105 [1281, 3353], respectively (Figure 3.12). For the remainder of this chapter, we will discard the strong and ultrastrong versions, and will refer to Bayes$_{\text{Weak}}$-dN simply as Bayes-dN.

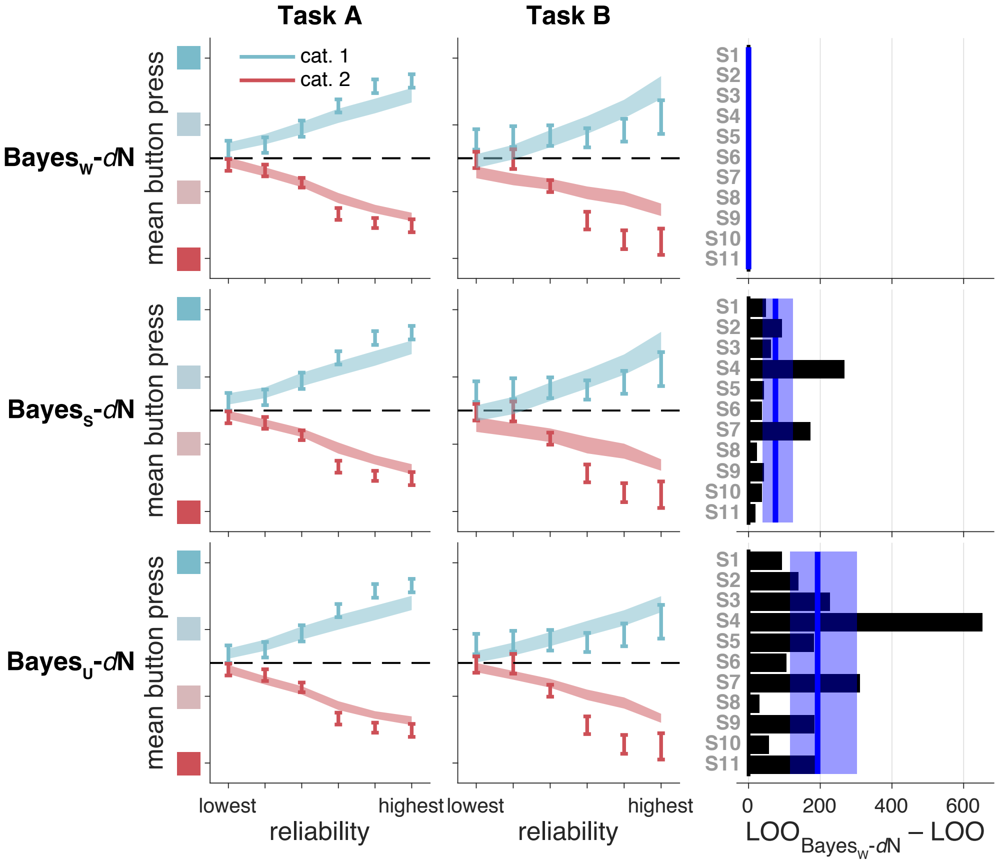

Figure 3.12 Model fits and model comparison for three strengths of the Bayesian model, as in Figure 3.10.

Weighted average of precision and perceived probability of being correct. A recent paper (Navajas et al., 2017) proposed that confidence is a weighted average of a function of variance, such as $\frac{1}{\sigma^2}$, and the perceived probability of being correct (incidentally, under a non-Bayesian decision rule). We tested such a model (using a Bayesian decision rule), which fits better than Fixed by 3059 [758, 5528] but still underperforms Lin by 3478 [2211, 5020].

Separate fits to Tasks A and B. In order to determine whether model rankings were primarily due to differences in one of the two tasks, we fit our models to each task individually.

In Task A, Quad fits better than Bayes-dN by 581 [278, 938], and better than Fixed by 3534 [2529, 4552] (Figure 3.14 and Table 3.1). In Task B, Quad fits better than Bayes-dN by 978 [406, 1756] and fits better than Fixed by 3234 [2099, 4390] (Figure 3.15 and Table 3.2).

Fits to category choice data only. In order to see whether our results were peculiar to combined category and confidence responses, we fit our models to the category choices only. Lin fits better than Bayes-dN by 595 [311, 927] and fits better than Fixed by 1690 [976, 2534] (Figure 3.16 and Table 3.3).

Fits to Task B only, with noise parameters fitted from Task A. To confirm that the fitted values of sensory uncertainty in the probabilistic models are meaningful, we treated Task A as an independent experiment to measure subjects' sensory noise. The category choice data from Task A can be used to determine the four uncertainty parameters. We fit Fixed with a decision boundary of 0° (equivalent to a Bayesian choice model with no prior), using maximum likelihood estimation. We fixed these parameters and used them to fit our models to Task B category and confidence responses. Lin fits better than Bayes-dN by 1773 [451, 2845] and fits better than Fixed by 5016 [3090, 6727] (Figure 3.17 and Table 3.4).

Separate category and confidence responses (experiment 2). There has been some recent debate as to whether it is more appropriate to collect choice and confidence with a single motor response (as described above) or with separate responses (Kiani et al., 2014; Navajas et al., 2016; Sanders et al., 2016; Wilimzig et al., 2008). Aitchison et al. (2015) found that confidence appears more Bayesian when subjects use separate responses. To confirm this, we ran a second experiment in which subjects chose a category by pressing one of two buttons, then reported confidence by pressing one of four buttons. Aitchison et al. (2015) also provided correctness feedback on every trial; in order to ensure that we could compare our results to theirs, we also provided correctness feedback in this experiment, even though this manipulation was not of primary interest. After fitting our core models, our results did not differ substantially from experiment 1: Lin fits better than Bayes-dN by 396 [186, 622] and fits better than Fixed by 2095 [1344, 2889] (Figure 3.18 and Table 3.5).

Task B only (experiment 3). It is possible that subjects behave suboptimally when they have to do multiple tasks in a session; in other words, perhaps one task "corrupts" the other. To explore this possibility, we ran an an experiment in which subjects completed Task B only. Quad fits better than Bayes-dN by 1361 [777, 2022] and fits better than Fixed by 7326 [4905, 9955] (Figure 3.19 and Table 3.6). In experiments 2 and 3, subjects only saw drifting Gabors; we did not use ellipses.

We also fit only the choice data, and found that Lin fits about as well as Bayes-dN, with summed LOO differences of 117 [-76, 436] and fits better than Fixed by 1084 [619, 1675] (Figure 3.20 and Table 3.7). This approximately replicates our previously published results (Qamar et al., 2013).

All model groupings. Below, we present model comparison results for all models, according to the groupings described in Section 3.2.4.7 (Figures 3.13 to 3.20 and Tables 3.1 to 3.7). We also present fits for our remaining core models (Figures 3.21 to 3.25); fits for Quad were shown in Figure 3.8.

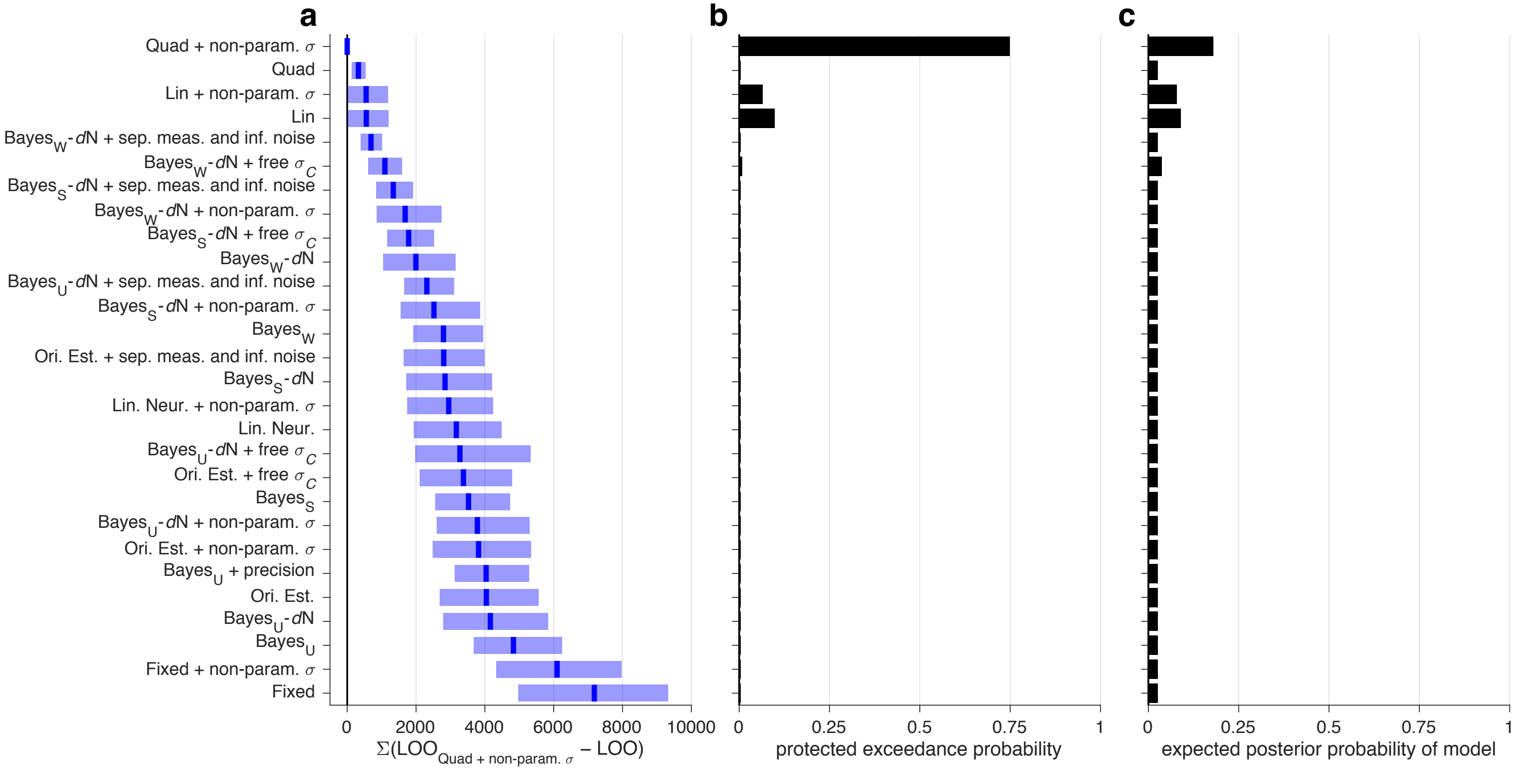

Figure 3.13 Model comparison, experiment 1. Models were fit jointly to Task A and B category and confidence responses. (**a**) Medians and 95% CI of bootstrapped sums of LOO differences, relative to the best model. Higher values indicate worse fits. (**b**) The protected exceedance probability, i.e., the posterior probability that a model occurs more frequently than the others (Rigoux et al., 2014). (**c**) The expected posterior probability that a model generated the data of a randomly chosen subject (Stephan et al., 2009). Note that due to the large number of models here, we do not including a cross comparison table like Tables 3.1 to 3.7.

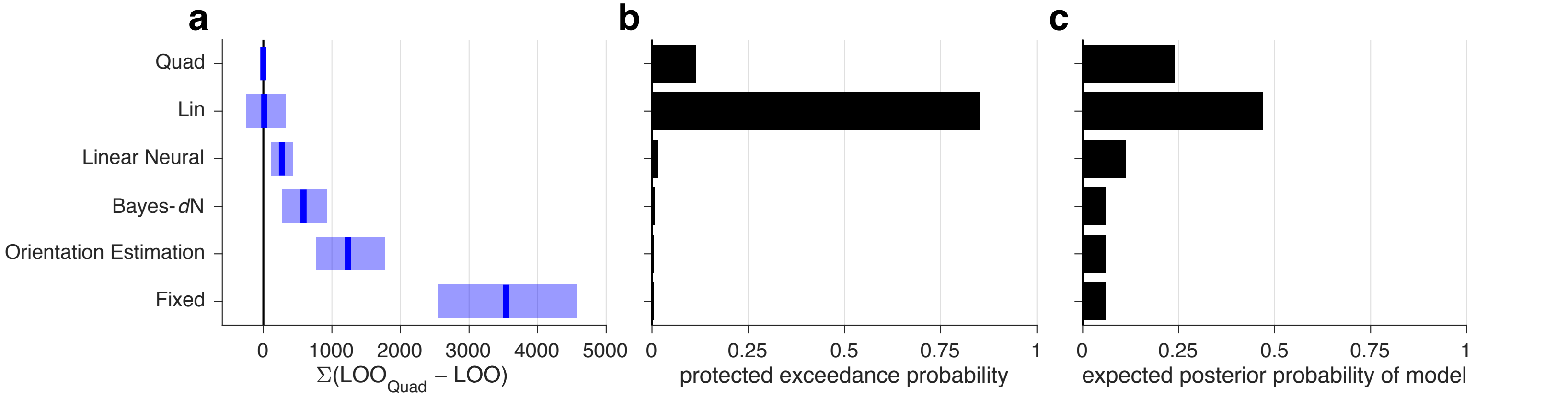

Figure 3.14 Model comparison, experiment 1. Models were fit to Task A category and confidence responses. See Figure 3.13 caption.

		12 pars. Fixed	13 pars. Bayes-dN	12 pars. Ori. Est.	13 pars. Lin. Neur.	16 pars. 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-dN	-2956 $[-3723, -2163]$				

Table 3.1 Cross comparison of all models in Figure 3.14. 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.

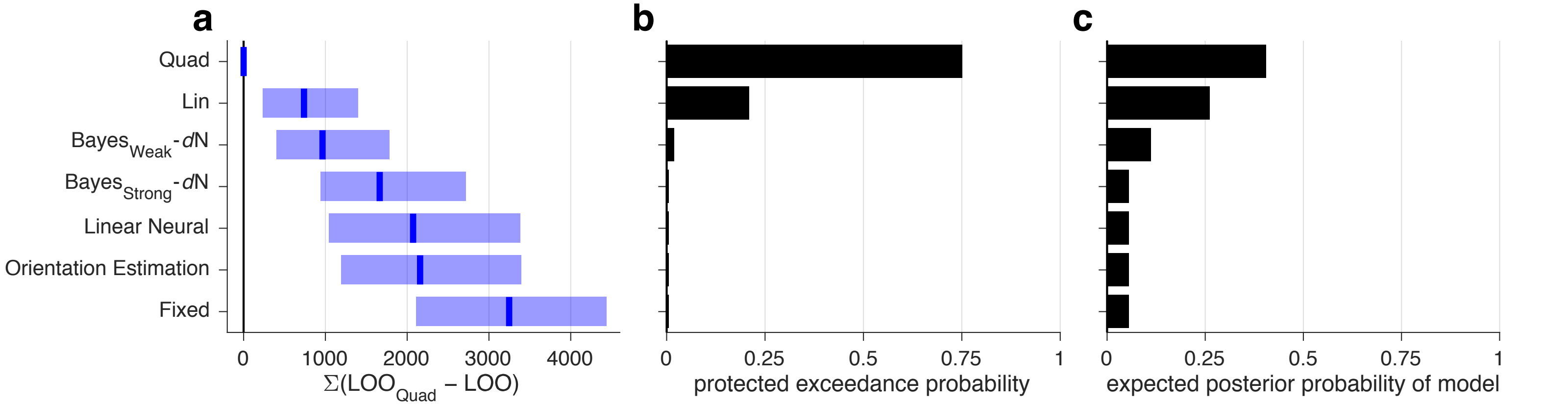

Figure 3.15 Model comparison, experiment 1. Models were fit to Task B category and confidence responses. See Figure 3.13 caption.

		15 pars.	13 pars.	16 pars.	15 pars.	16 pars.	22 pars.
		Fixed	Bayes$_S$-dN	Bayes$_W$-dN	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$_W$-dN	-2230 [$-3239, -1307$]	-691 [$-1082, -390$]				
13 pars.	Bayes$_S$-dN	-1534 [$-2425, -634$]					

Table 3.2 Cross comparison of all models in Figure 3.15. See Table 3.1 caption.

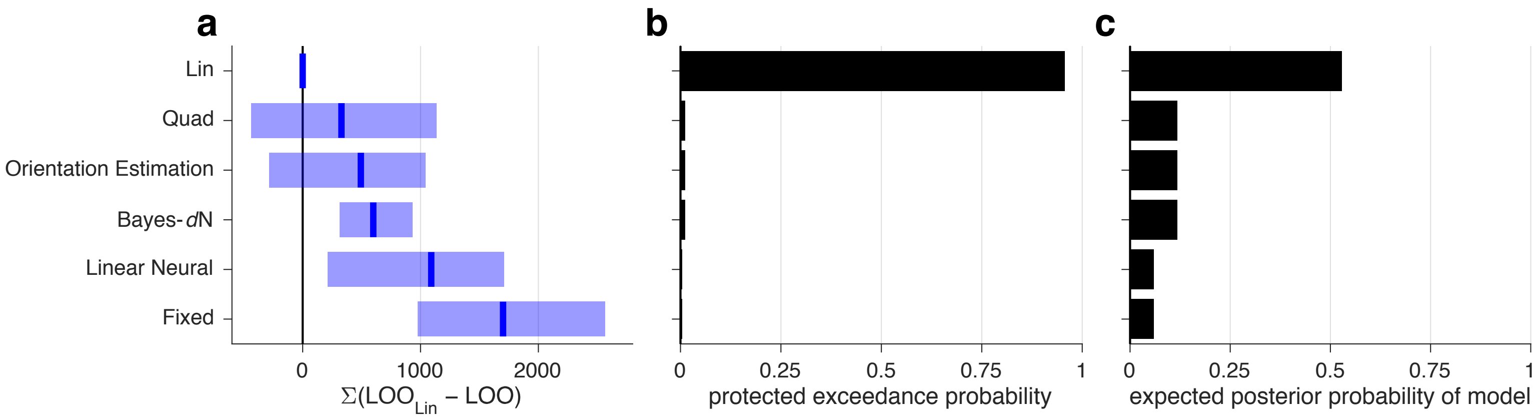

Figure 3.16 Model comparison, experiment 1. Models were fit jointly to Task A and B category choices. See Figure 3.13 caption.

		8 pars.	9 pars.	8 pars.	9 pars.	10 pars.
		Fixed	Bayes-dN	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-dN	-1095 $[-1657, -629]$				

Table 3.3 Cross comparison of all models in Figure 3.16. See Table 3.1 caption.

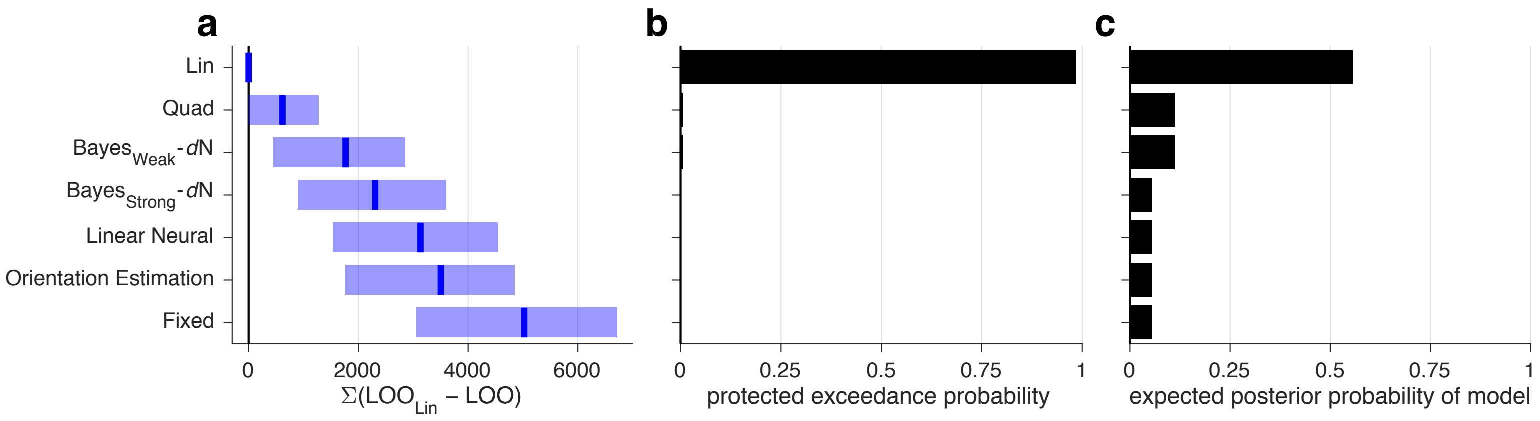

Figure 3.17 Model comparison, experiment 1. Noise parameters were fit to Task A category choices and then fixed during the fitting of Task B category and confidence responses. See Figure 3.13 caption.

		15 pars.	13 pars.	16 pars.	15 pars.	16 pars.	22 pars.
		Fixed	Bayes$_{\text{S}}$-dN	Bayes$_{\text{W}}$-dN	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$_{\text{W}}$-dN	-3257 $[-3965, -2494]$	-533 $[-920, -283]$				
13 pars.	Bayes$_{\text{S}}$-dN	-2704 $[-3351, -2027]$					

Table 3.4 Cross comparison of all models in Figure 3.17. See Table 3.1 caption.

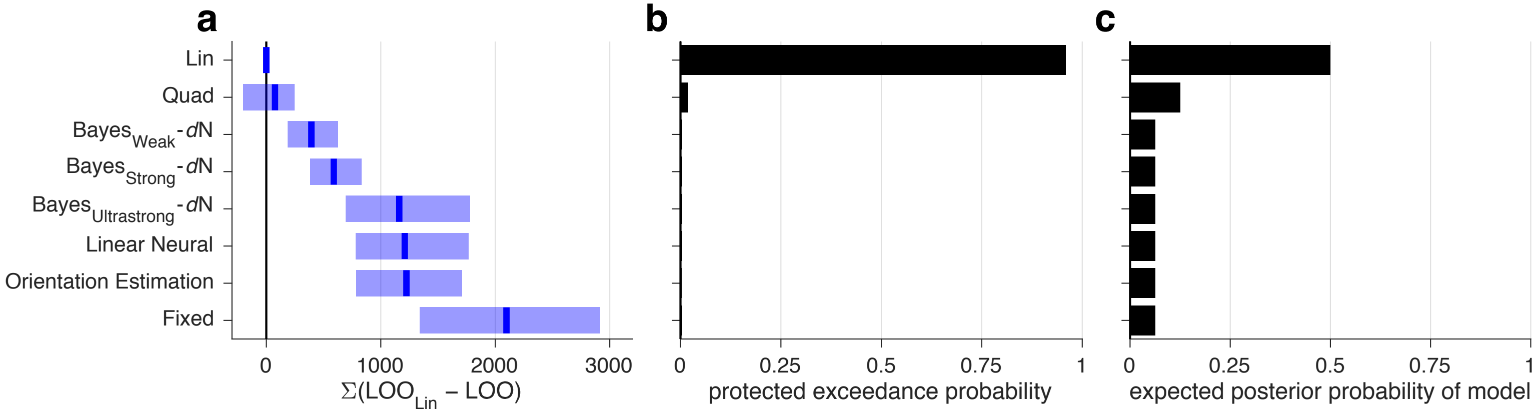

Figure 3.18 Model comparison, experiment 2. Models were fit jointly to Task A and B category and confidence responses. See Figure 3.13 caption.

		19 pars.	13 pars.	17 pars.	20 pars.	19 pars.	20 pars.	30 pars.
		Fixed	Bayes_{U}-dN	Bayes_{S}-dN	Bayes_{W}-dN	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_{W}-dN	-1678 $[-2542, -1007]$	-767 $[-1262, -386]$	-190 $[-363, -82]$				
17 pars.	Bayes_{S}-dN	-1490 $[-2210, -886]$	-565 $[-1032, -266]$					
13 pars.	Bayes_{U}-dN	-907 $[-1572, -365]$						

Table 3.5 Cross comparison of all models in Figure 3.18. See Table 3.1 caption.

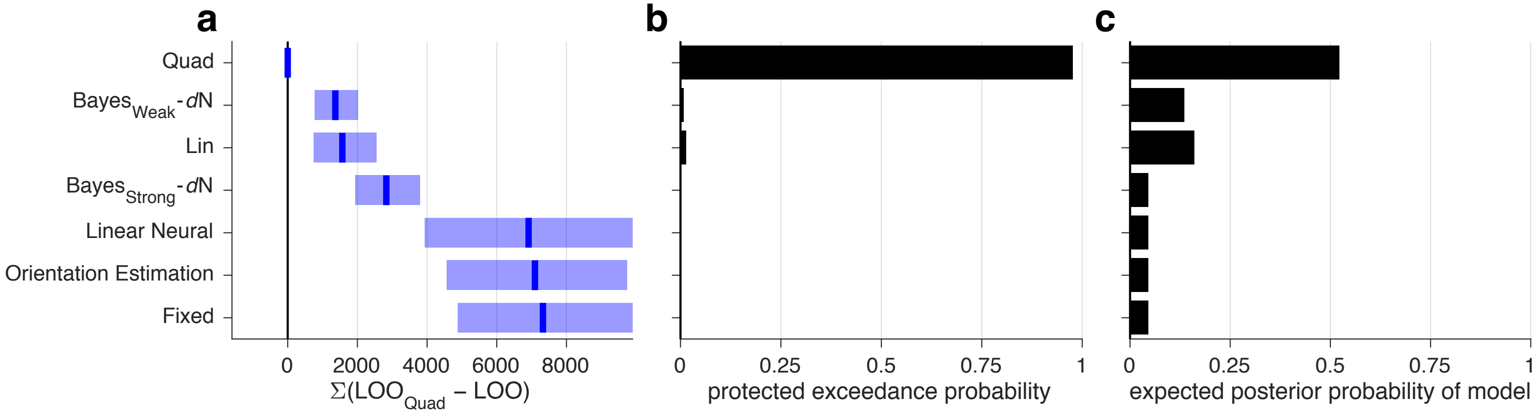

Figure 3.19 Model comparison, experiment 3. Models were fit to Task B category and confidence responses. See Figure 3.13 caption.

		15 pars.	13 pars.	16 pars.	15 pars.	16 pars.	22 pars.
		Fixed	Bayes$_{\text{S}}$-dN	Bayes$_{\text{W}}$-dN	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$_{\text{W}}$-dN	-5967 $[-8702, -3369]$	-1454 $[-2179, -835]$				
13 pars.	Bayes$_{\text{S}}$-dN	-4505 $[-7282, -1816]$					

Table 3.6 Cross comparison of all models in Figure 3.19. See Table 3.1 caption.

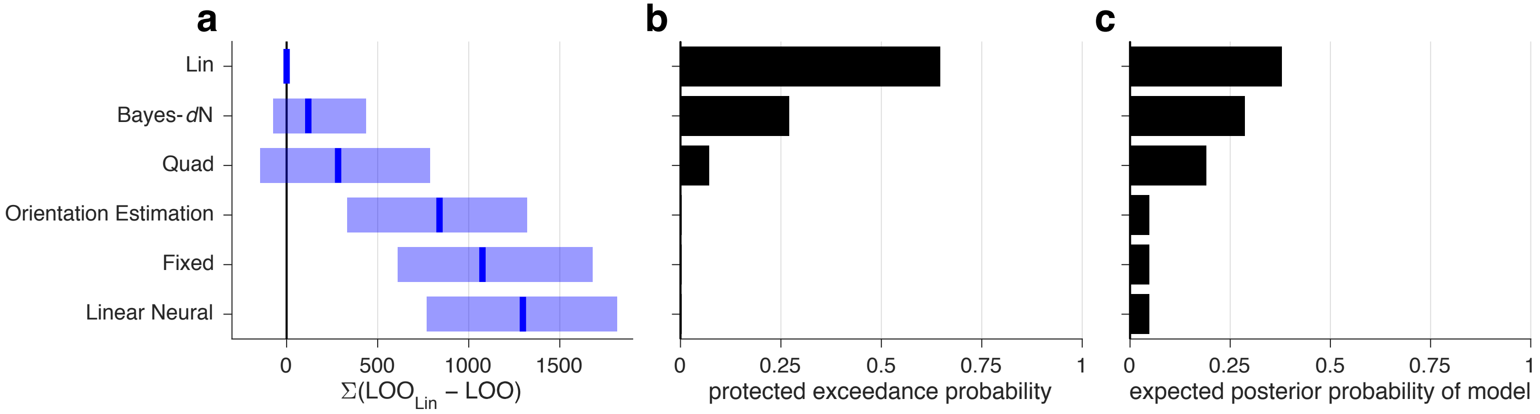

Figure 3.20 Model comparison, experiment 3. Models were fit to Task B category choices. See Figure 3.13 caption.

		7 pars.	8 pars.	7 pars.	8 pars.	8 pars.
		Fixed	Bayes-dN	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-dN	-964 $[-1290, -663]$				

Table 3.7 Cross comparison of all models in Figure 3.20. See Table 3.1 caption.

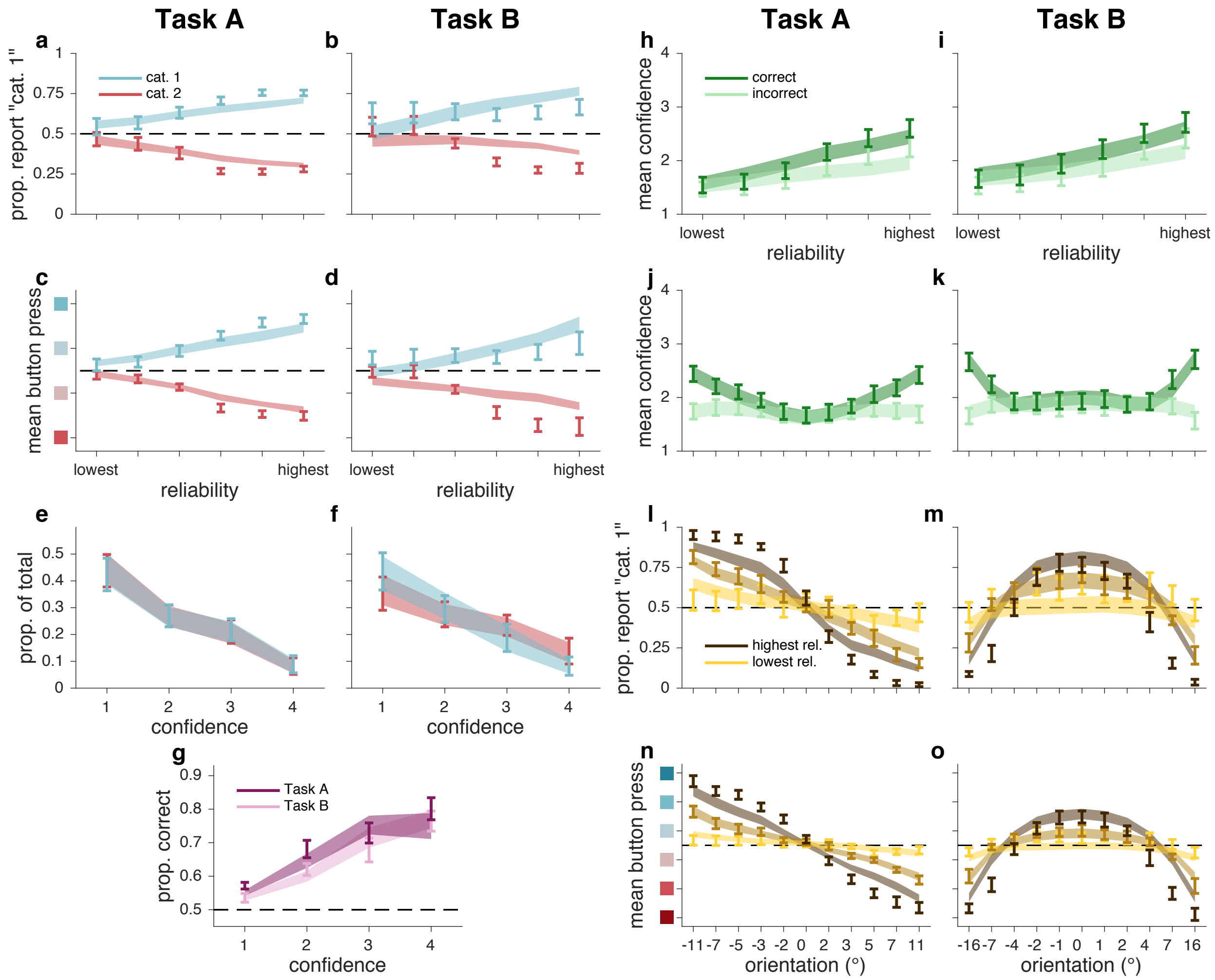

Figure 3.21 Bayes-dN fits, as in Figure 3.8

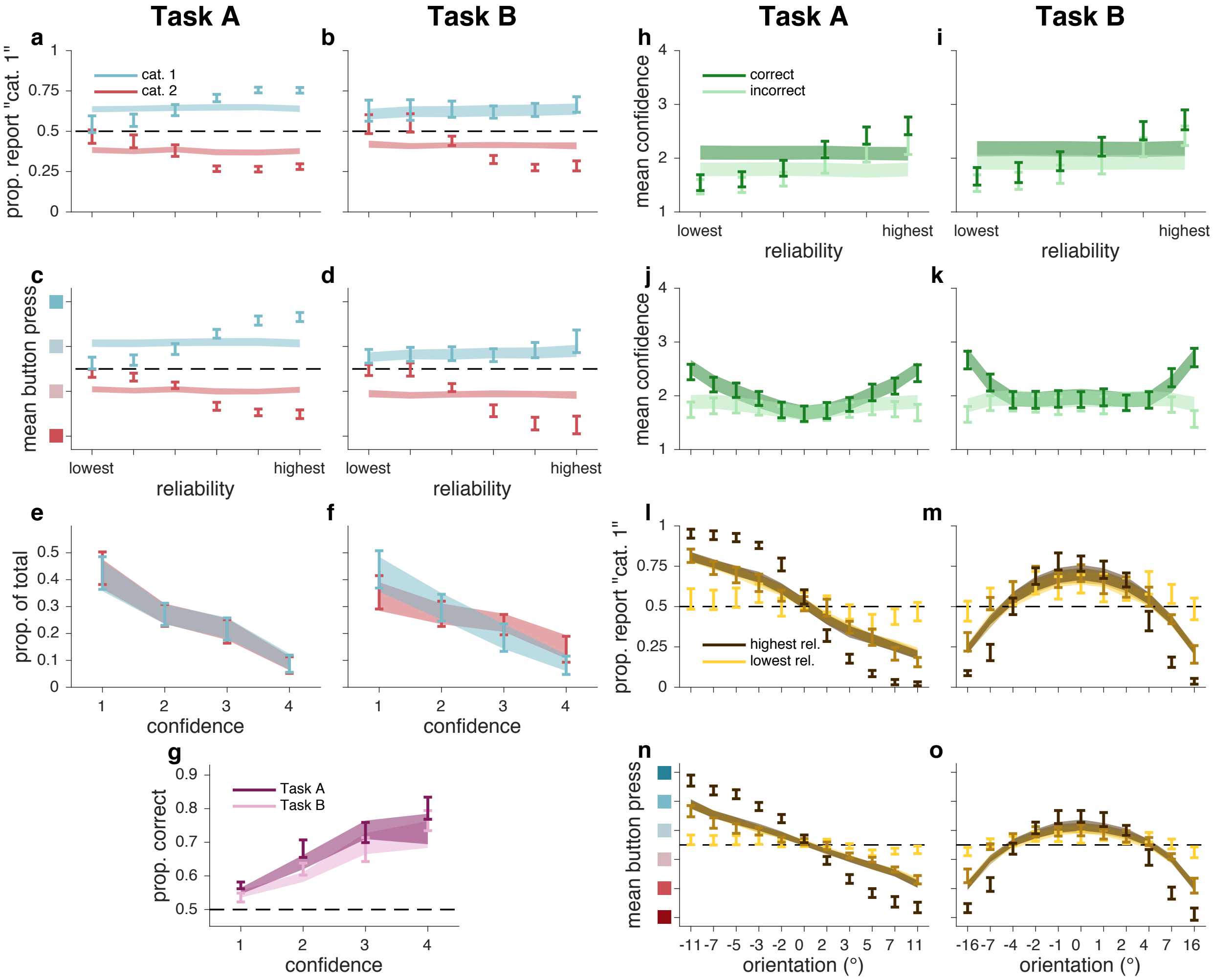

Figure 3.22 Fixed fits, as in Figure 3.8.

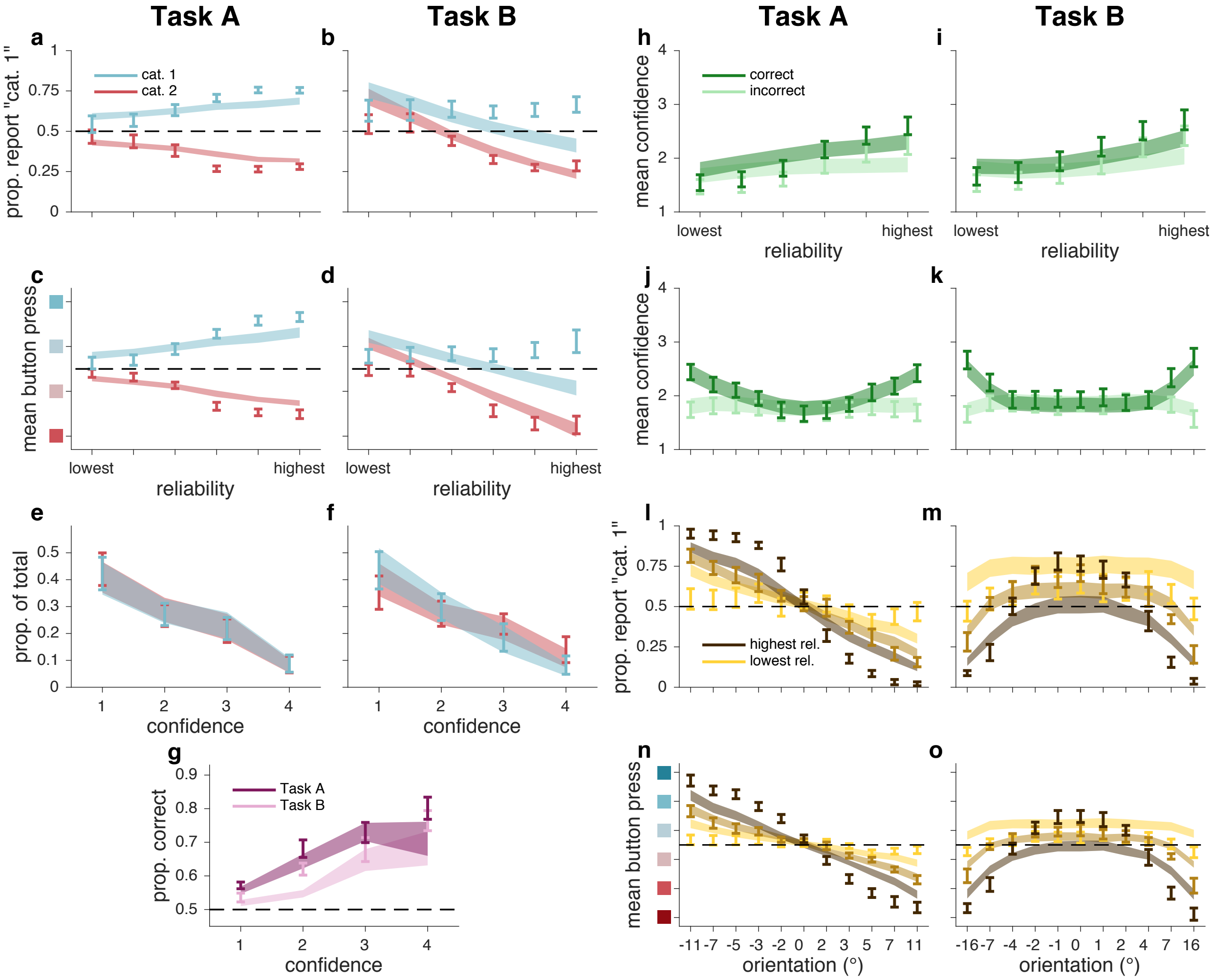

Figure 3.23 Orientation Estimation fits, as in Figure 3.8.

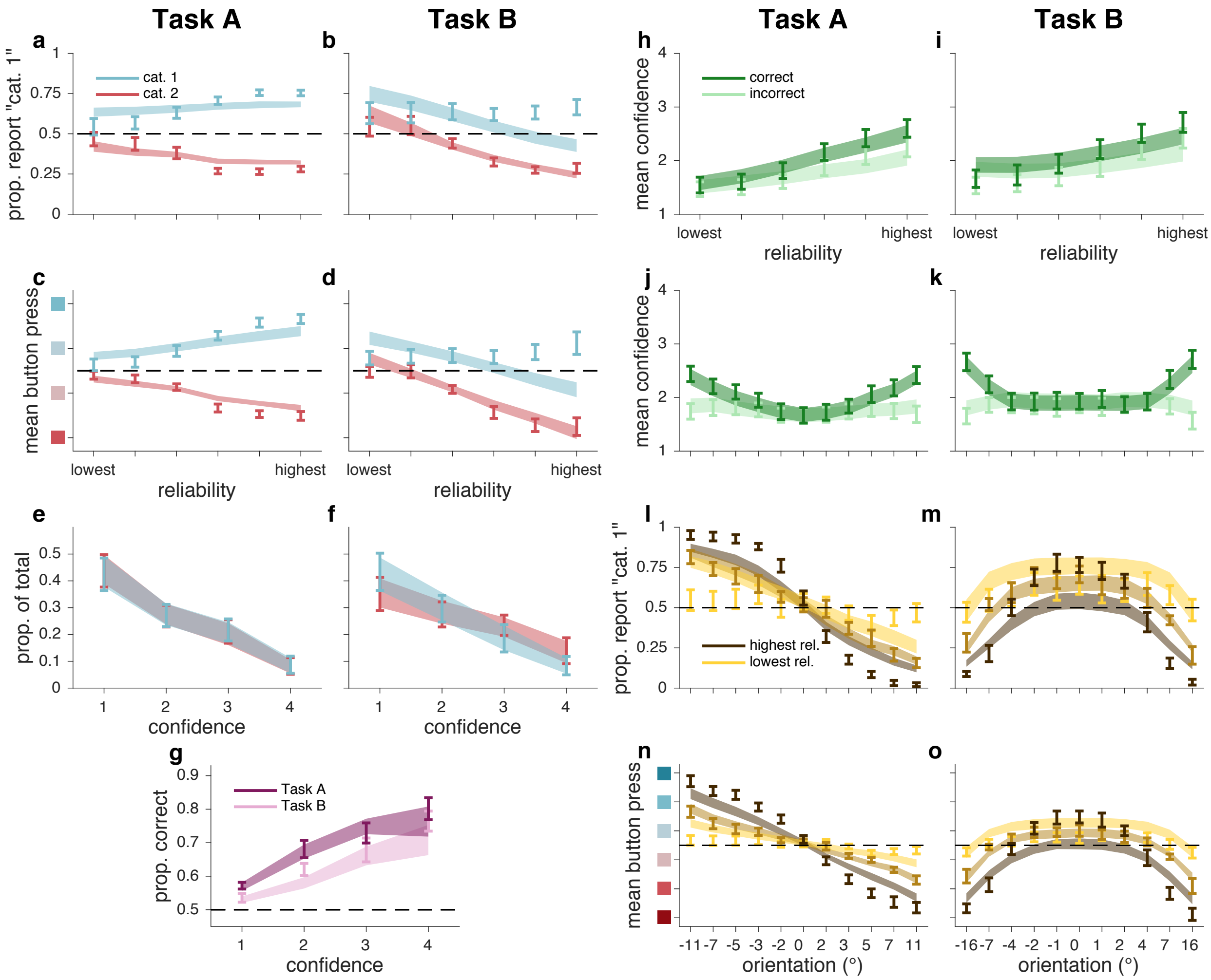

Figure 3.24 Linear Neural fits, as in Figure 3.8.

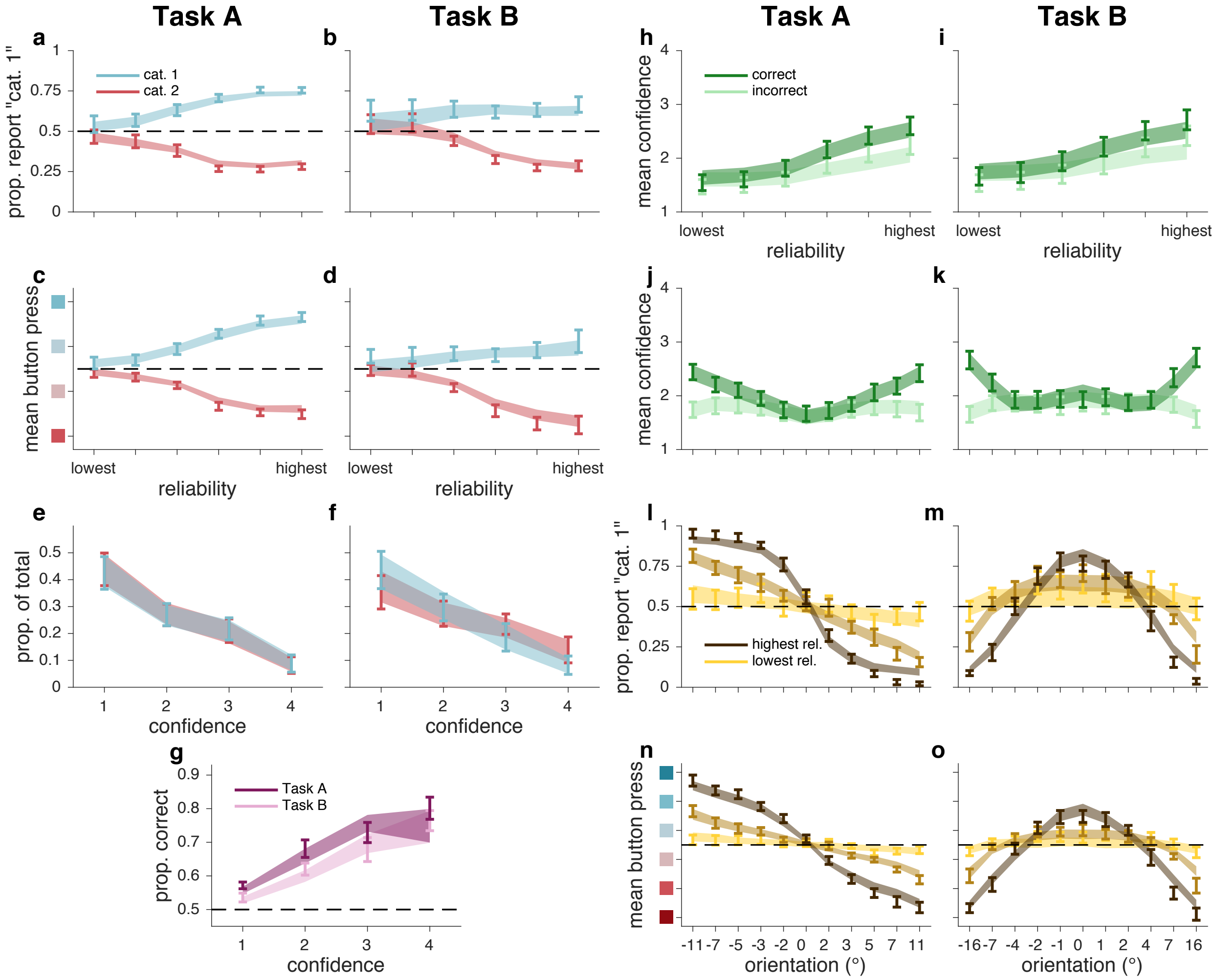

Figure 3.25 Lin fits, as in Figure 3.8.

3.3.3 Effect of stimulus type on model comparison results

In experiment 1, since some subjects only saw Gabors and some only saw ellipses, we used Spearman's rank correlation coefficient to measure the similarity of the two groups' model rankings. Spearman's rank correlation coefficient between Gabor and ellipse subjects for the summed LOO scores of the model groupings in Figure 3.11 and Figure 3.13 was 0.952 and 0.944, respectively (a value of 1 would indicate identical rankings). In both model groupings, the identities of the lowest- and highest-ranked models were the same for both Gabor and ellipse subjects. This indicates that the choice of stimulus type did not have a systematic effect on model rankings.

3.3.4 Model comparison metric analysis

We determined that our results were not dependent on our choice of metric. We computed AIC, BIC, AICc, WAIC, and LOO for all models in the 8 model groupings, multiplying the information criteria by $-\frac{1}{2}$ to match the scale of LOO (Section 3.2.4.7). 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 rank correlation coefficient 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.

3.3.5 Model recovery

We performed a model recovery analysis (van den Berg et al., 2014) to test our ability to distinguish our 6 core models, as well as the 2 stronger versions of the Bayesian model. We generated synthetic datasets from each of the 8 models 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 3.2.4.6) 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, using maximum likelihood estimation (MLE) of parameters by an interior-point constrained optimization (MATLAB's *fmincon*), and computed AIC scores from the resulting fits.

We found that the true generating model was the best-fitting model, on average, in all cases (Figure 3.26). 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.

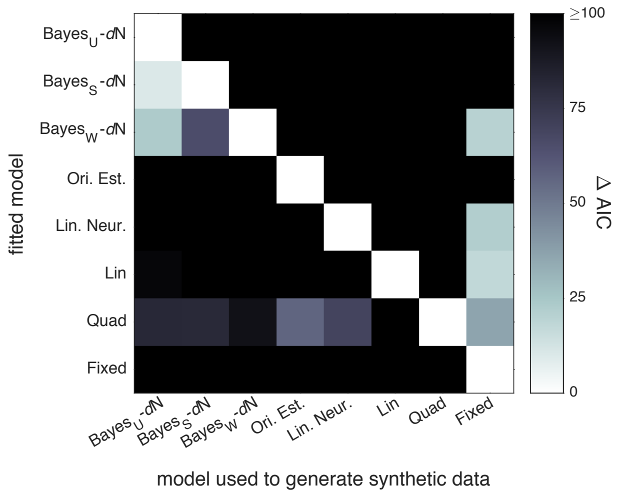

Figure 3.26 Model recovery analysis. Shade represents the difference between the mean AIC score (across datasets) for each fitted model and for the one with the lowest mean AIC score. White squares indicate the model that had the lowest mean AIC score when fitted to data generated from each model. The squares on the diagonal indicate that the true generating model was the best-fitting model, on average, in all cases.

Ideally, we would have evaluated our model recovery fits using LOO, as we evaluated the fits to human data. However, LOO can only be obtained when fitting with MCMC sampling, which takes orders of magnitudes longer than fitting with MLE. It would be impossible to fit all 352 synthetic datasets in a short amount of time using the same procedure and sampling quality standards described in Section 3.2.4.6 (i.e., a large number of samples, across multiple converged chains). Furthermore, we do not believe that our model recovery is dependent on how the models are fit and the fits are evaluated; we found that AIC and LOO scores gave us near-identical model rankings for data from real subjects (Section 3.3.4).

3.4 Discussion

We carried out a strong test of whether human confidence reports are Bayesian, using overlapping categories (Palminteri et al., 2017), withholding feedback on testing trials, and varying experimental components such as task, stimulus type, and stimulus reliability (Maloney and Mamassian, 2009). We used model comparison to investigate the computational underpinnings of confidence, fitting a total of 75 models from 6 distinct model families.

Our first finding is that, like the optimal observer, subjects use knowledge of their sensory uncertainty when reporting confidence in a categorical decision; models in which the observer ignores their sensory uncertainty provide a poor fit to the data (Figure 3.9). Our second finding is that subjects do not appear to use knowledge of their sensory uncertainty in a way that is fully consistent with the Bayesian confidence hypothesis. Instead, heuristic models that approximate Bayesian computation—but do not compute a posterior probability over category—outperform the Bayesian models in two tasks (Figure 3.10, compare top row to bottom two rows). This result continued to hold after we relaxed assumptions about the relationship between reliability and noise, and about the subject's knowledge of the generating model. We accounted for the fact that our models had different amounts of flexibility by using a wide array of model comparison metrics and by showing that our models are meaningfully distinguishable.

Our conclusions differ from those of some recent experimental findings. Like the present study, Aitchison et al. (2015) found evidence that confidence reports may emerge from heuristic computations. However, they sampled stimuli from only a small region of their two-dimensional space, where model predictions may not vary greatly. Therefore, their stimulus set did not allow for the models to be strongly distinguished. Furthermore, although they tested for *Bayesian* computation, they did not test for *probabilistic* computation (whether observers take sensory uncertainty into account on a trial-to-trial basis (Ma, 2012)) as we do here. Such a test requires that the experimenter vary the reliability, not only the value, of the stimulus feature of interest.

Sanders et al. (2016) reported that confidence has a “statistical” nature. However, their experiment was unable to determine whether confidence is Bayesian or not (Ma and Jazayeri, 2014), because the stimuli varied along only one dimension. Aitchison et al. (2015) note that, to distinguish models of confidence, the experimenter must use stimuli that are characterized by two dimensions (e.g., contrast and orientation). This is because, when fitting models that map from an internal variable to an integer confidence rating, it is impossible to distinguish between two internal variables that are monotonically related (in the case of Sanders et al. (2016), the measurement and the posterior probability of being correct). Therefore, the only alternative model proposed by Sanders et al. (2016) is based on reaction time, rather than on the presented stimuli.

Navajas et al. (2017) suggested that confidence reports are best described as a weighted average of precision and the probability of being correct. However, their model uses the estimated probability of being correct under a non-Bayesian decision rule (Section 2.6.2). They did not show the fit of a Bayesian model, and therefore their study does not constitute a true test of whether confidence is Bayesian. Here, we tested and rejected the hypothesis that confidence is a weighted average of precision and the posterior probability of being correct under a Bayesian decision rule.

Our results raise general issues about the status of Bayesian models as descriptions of behavior. First, because it is impossible to exhaustively test all models that might be considered “Bayesian,” we cannot rule out the entire class of models. However, we have tried to alleviate this issue as much as possible by testing a large number of Bayesian models—far more than the number of Bayesian and non-Bayesian models tested in other studies of confidence. Second, Bayesian models are often held in favor for their generalizability; one can determine the performance-maximizing strategy for any task. Although generalizability indeed makes Bayesian models attractive and powerful, we do not believe that this property should override a bad fit.

In the next chapter, we will test a different manipulation of sensory uncertainty.

Chapter 4

Human confidence reports under top-down stimulus uncertainty

4.1 Introduction

In the previous chapter, we tested whether confidence ratings were Bayesian when stimulus uncertainty came from bottom-up factors, external to the observer. In particular, we adjusted uncertainty by manipulating stimulus contrast or elongation. Uncertainty, however, can originate not only from the external world but also from one's internal state.

Attention is a critical internal state variable that governs the uncertainty of visual representations (Carrasco, 2011; Reynolds and Chelazzi, 2004); it modulates basic perceptual properties like contrast sensitivity (Carrasco et al., 2000; Lu and Dosher, 1998) and spatial resolution (Anton-Erxleben and Carrasco, 2013). Surprisingly, it has been suggested that, unlike for external sources of uncertainty, people fail to take attention into account during perceptual decision-making (Morales et al., 2015; Rahnev et al., 2011, 2012a), leading to inaccurate decisions and overconfidence—a risk in attentionally demanding situations like driving a car. However, this proposal has never been directly tested using formal model comparison.

The work presented in this Chapter is similar to the work presented in Chapter 3, Experiment 1, with three major changes. First, we induce stimulus uncertainty by manipulating subjects' attention levels rather than stimulus reliability. Second, subjects only perform Task B, rather than both Tasks A and B. Third, while we tested dozens of models in Chapter 3, here we test only a handful of representative models.

As in Chapter 3, we find that the BCH qualitatively describes behavior, and fits much better than the Fixed model, in which observers ignore their uncertainty. Unlike in Chapter 3, however, we are unable to distinguish the Bayesian model from heuristic models that take sensory uncertainty into account in a non-Bayesian way.

4.2 Methods

4.2.1 Experiment

Observers completed the Task B categorization task described in Section 3.2.1, except that we manipulated stimulus uncertainty by cuing observers to pay more or less attention to a stimulus. This required us to make several modifications to the experiment. On each trial, four stimuli were briefly presented on each trial, and a response cue indicated which stimulus to report. Preceding the stimulus presentation, we manipulated voluntary (i.e., endogenous) attention on a trial-to-trial basis using a spatial cue that pointed to either one stimulus location (*valid* condition: the response cue matched the cue, 66.7% of trials; and *invalid* condition: it did not match, 16.7% of trials) or all four locations (*neutral* condition: 16.7% of trials) (Figure 4.1b). Twelve subjects participated, with about 2000 trials per subject.

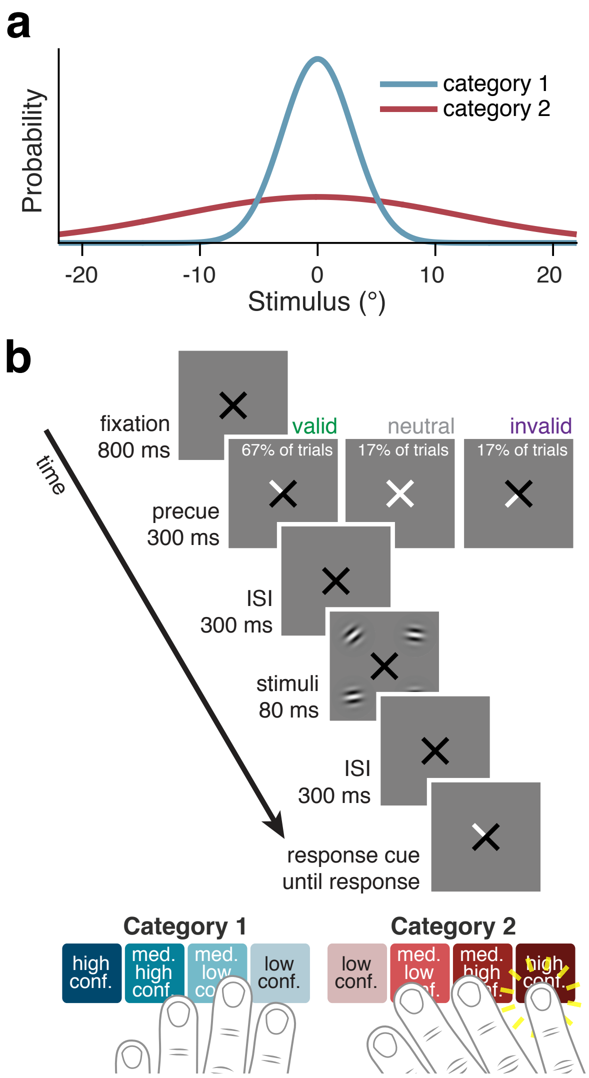

Figure 4.1 Stimuli and task. (**a**) Stimulus orientation distributions for each category. (**b**) Trial sequence. Cue validity, the likelihood that a precue to one quadrant would match the response cue, was 80%.

4.2.1.1 Subjects

Twelve subjects (7 female, 5 male), aged 18–25 years, participated in the experiment. These subjects came from an original set of 28 subjects who completed at least one session. The remaining subjects did not complete the experiment, either because they were excluded on the basis of their staircase performance (Section 4.2.1.3) or because they chose to stop participating before all sessions were completed. Subjects received \$10 per 40–60 minute session, plus a completion bonus of \$25. The experiments were approved by the University Committee on Activities Involving Human Subjects of New York University. Informed consent was given by each subject before the experiment. All subjects were naïve to the purpose of the experiment. No subjects were fellow scientists.

4.2.1.2 Apparatus and stimuli

Apparatus Subjects were seated in a dark room, at a viewing distance of 57 cm from the screen, with their chin in a chinrest. Stimuli were presented on a gamma-corrected 100 Hz, 21-inch display (Model Sony GDM-5402). The display was connected to a 2010 iMac running OS X 10.6.8 using MATLAB (Mathworks) with Psychophysics Toolbox 3 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997).

Stimuli The background was mid-level gray (60 cd/m^2). Stimuli consisted of drifting Gabors with a spatial frequency of 0.8 cycles per degree, a speed of 6 cycles/s, a Gaussian envelope with a s.d. of 0.8 degrees of visual angle (dva), and a randomized starting phase. In category training, the stimuli were positioned at fixation, and the central fixation cross was a black "$+$" subtending 1.2 dva in diameter. In all other blocks, one stimulus was positioned in each of the four quadrants of the screen, at 45, 135, 225, and 315 degrees, 5 dva from fixation, and the fixation cross was a black "$\times$" with each arm pointing to a quadrant. One or more of the arms turned white to provide a precue or response cue (Figure 4.1b). Stimulus contrast depended on the block type.

Categories Stimulus categories were as described in Section 3.2.1 for Task B.

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

4.2.1.3 Procedure

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

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

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

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

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

Instructions *First staircase session.* Before the first category training block, we provided subjects with a printed graphic similar to Figure 4.1a, explained how the stimuli were generated from distributions, and explained the category training procedure. We also explained that trials would only proceed when the subject maintained fixation. Before the category/attention training block, we explained the attention task using an onscreen graphic that explained the cuing procedure and a printed graphic that illustrated cue validity. We also explained the requirement to maintain fixation from the precue until the response cue and the consequences of breaking fixation. Before the first category/attention testing-with-staircase block, we explained that the stimulus presentation time would be shorter and that the contrast of the stimuli would vary.

First test session. Before the confidence/attention training block, we explained two changes to the experiment. First, we told subjects that they would be reporting category choice and confidence simultaneously. We provided a printed graphic similar to the buttons shown in Figure 4.1b, showing the eight buttons representing category choice and confidence level, the latter on a 4-point scale. The confidence levels were labeled as "very high," "somewhat high," "somewhat low," and "very low." All printed graphics were visible to subjects throughout the experiment. Second, we told subjects that contrast would be fixed (rather than variable) for the remainder of the experiment, in all blocks.

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

Category training To ensure that subjects knew the stimulus distributions well, we gave them extensive category training with trial-to-trial correctness feedback and foveal stimulus presentation to reduce orientation uncertainty. Each trial proceeded as follows: 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 and displayed as a drifting Gabor. 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 correctness feedback was displayed for 1.1 s, e.g., "You said Category 1. Correct!" The fixation cross then reappeared.

In staircase sessions, the stimulus contrast was 35%. In test sessions, the contrast matched the subject-specific levels chosen for testing blocks, in order to minimize obvious changes between training and testing blocks. Each category training block had 2 sets of 36 trials (72 total). At the end of the block, subjects were shown the percentage of trials that they had correctly categorized.

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

Category/attention testing-with-staircase The purpose of this block was to determine the stimulus contrast for each subject that would be used in the test sessions. The trial procedure was identical to that of category/attention training, except that stimulus presentation time was 80 ms (instead of 300 ms) and stimulus contrast varied. We used an adaptive staircase procedure to determine the stimulus contrast on each trial and estimate psychometric functions for performance accuracy as a function of log contrast. Separate staircases were used for valid, neutral, and invalid conditions. We used Luigi Acerbi's MATLAB implementation (github.com/lacerbi/psybayes) of the PSI method by Kontsevich and Tyler (Kontsevich and Tyler, 1999), extended to include the lapse rate (Prins, 2012). The method generates a posterior distribution over three parameters of the psychometric function: threshold μ, slope σ, and lapse rate λ. On each trial, it selects a stimulus intensity that maximizes the expected information gain by completion of the trial (note in Figure 4.2 that the selected trials are most numerous where the slopes of the psychometric curves are highest). μ (log contrast units) ranged from -6.5 to 0 and had a Gaussian prior distribution with mean -2 and s.d. 1.2. $\log \sigma$ ranged from -3 to 0, and had a uniform prior distribution across the range. λ ranged from 0.15 (because the maximum accuracy in the task was slightly below $1 - 0.15$) to 0.5, and had a Beta prior distribution with shape parameters $\alpha = 20$ and $\beta = 39$. Each block had 4 sets of 36 trials (144 total). At the end of the block, subjects were shown the percentage of trials that they had correctly categorized.

Subject and contrast selection After each subject's final staircase session, we plotted and visually inspected the mean and s.d. of the posterior over the 3 (valid, neutral, and invalid) estimated psychometric functions (an example is shown in Figure 4.2). A subject was considered eligible for the remainder of the study if there existed a contrast level at which the mean minus the s.d. of the posterior over invalid psychometric functions was above chance, and the mean minus the s.d. of the posterior over valid psychometric functions was greater than the mean plus 1 s.d. of the posterior over invalid psychometric functions (for example, note that there is a range of values in Figure 4.2 for which the red shading does not overlap with the chance line or with the green shading). Subjects for which there were no suitable contrast levels did not continue the study. Each experimenter selected a log contrast for which the separation between valid, neutral, and invalid performance appeared, by visual inspection, to be maximal. We then took the average of these log contrast values.

We used this subject screening and contrast selection procedure because, in order to test our hypothesis, we needed uncertainty to depend on attention. This procedure increased the probability that uncertainty would vary between attention conditions in the final dataset. Selected contrasts ranged from 4% to 60% across subjects.

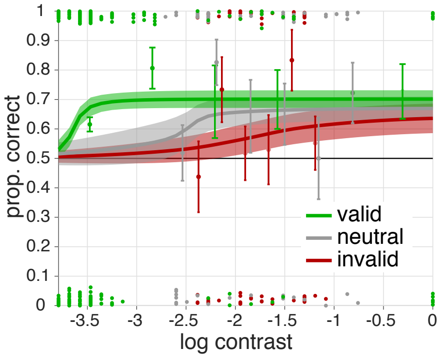

Figure 4.2 Example plot used to determine per-subject stimulus contrast. Each curve shows the mean ±1 s.d. of the posterior over psychometric functions for each attention condition. Error bars indicate the mean ±1 s.d. of the beta distribution over correctness within log contrast bins. A dot indicates one correct or incorrect trial, located respectively at the top or bottom of the plot, with vertical jitter. For this example subject, we selected a log contrast of -2.3 (i.e., a contrast of 10%).

Confidence/attention training To familiarize subjects with the button mappings for choice and confidence, they completed confidence/attention training. The trial procedure was identical to category/attention training, except subjects reported their confidence on each trial in addition to their category choice. 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. Feedback about their choice and confidence report was presented for 1.2 s after each trial, e.g., "You said category 2 with HIGH confidence." The stimulus contrast was specific to each subject, based on the staircase sessions. There were 30 trials per block.

Confidence/attention testing These were the main experimental blocks. The trial procedure (Figure 4.1b) was the same as in confidence/attention training blocks, but with no trial-to-trial feedback whatsoever. Each block had 4 sets of 36 trials (144 total). At the end of each block, subjects were required to take a break of at least 30 s. During the break, they were shown the percentage of trials that they had correctly categorized. Subjects were also shown a list of the top 10 block scores (across all subjects, indicated by initials). This was intended to motivate subjects to perform well, and to reassure them that their scores were normal, since it is rare to score above 75% on a block.

4.2.2 Modeling

We tested two sets of models: category choice and confidence models, and category choice-only models. Model parameters are described in Tables 4.1 and 4.2. The model referred to in Chapter 3 as Bayes-*d*N or Bayes is referred to as Bayesian in this chapter. Model specification and fitting procedures for this chapter were as described in Section 3.2.4, except for the following differences.

4.2.2.1 Model specification

Free We fit a Free model in which the observer compares the orientation measurement to a set of boundaries that vary nonparametrically (i.e., free of a parametric relationship with σ) across attention conditions. We used this model only for the purpose of obtaining estimates of the category decision boundary parameters. We fit free parameters $k_{4,\text{valid}}$, $k_{4,\text{neutral}}$, $k_{4,\text{invalid}}$, and used measurement boundaries $b_{4,\text{attention condition}} = k_{4,\text{attention condition}}$.

4.2.2.2 Model fitting

For each chain, we took 100,000 to 1,000,000 total samples (depending on model computational time) from the posterior distribution over parameters. We discarded the first third of the samples and kept 6,667 of the remaining samples, evenly spaced to reduce autocorrelation. All samples with log posteriors more than 40 below the maximum log posterior were discarded. Marginal probability distributions of the sample log likelihoods were visually checked for convergence across chains. In total we had 120 model and dataset combinations, with a median of 40,002 kept samples (interquartile range = 13,334).

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

Table 4.1 Parameters of category choice and confidence decision models. Note that we did not present a corresponding table like this for the models presented in Chapter 3, due to the large number of models.

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

Table 4.2 Parameters of category choice-only decision models. * indicates models that were used only for obtaining parameter estimates (Figure 4.8, Figure 4.7c), and not for model comparison. Note that we did not present a corresponding table like this for the models presented in Chapter 3, due to the large number of models.

4.3 Results

4.3.1 Descriptive statistics

Cue validity increased categorization accuracy [one-way repeated-measures ANOVA, $F(2,11) = 95.88$, $p < 10^{-10}$], with higher accuracy following valid cues [two-tailed paired t-test, $t(11) = 7.92$, $p < 10^{-5}$] and lower accuracy following invalid cues [$t(11) = 4.62$, $p < 10^{-3}$], relative to neutral cues (Figure 4.3a, left). This pattern confirms that attention increased orientation sensitivity (e.g., (Cameron et al., 2002; Lu and Dosher, 1998)). Attention also increased confidence ratings [$F(2,11) = 13.35$, $p < 10^{-3}$] and decreased reaction time [$F(2,11) = 28.76$, $p < 10^{-6}$], ruling out speed-accuracy tradeoffs as underlying the effect of attention on accuracy (Figure 4.3a).

Decision rules in this task are defined by how they map stimulus orientation and attention condition onto a response. We therefore plotted behavior as a function of these two variables. Overall performance was a “W”-shaped function of stimulus orientation (Figure 4.3b, left),

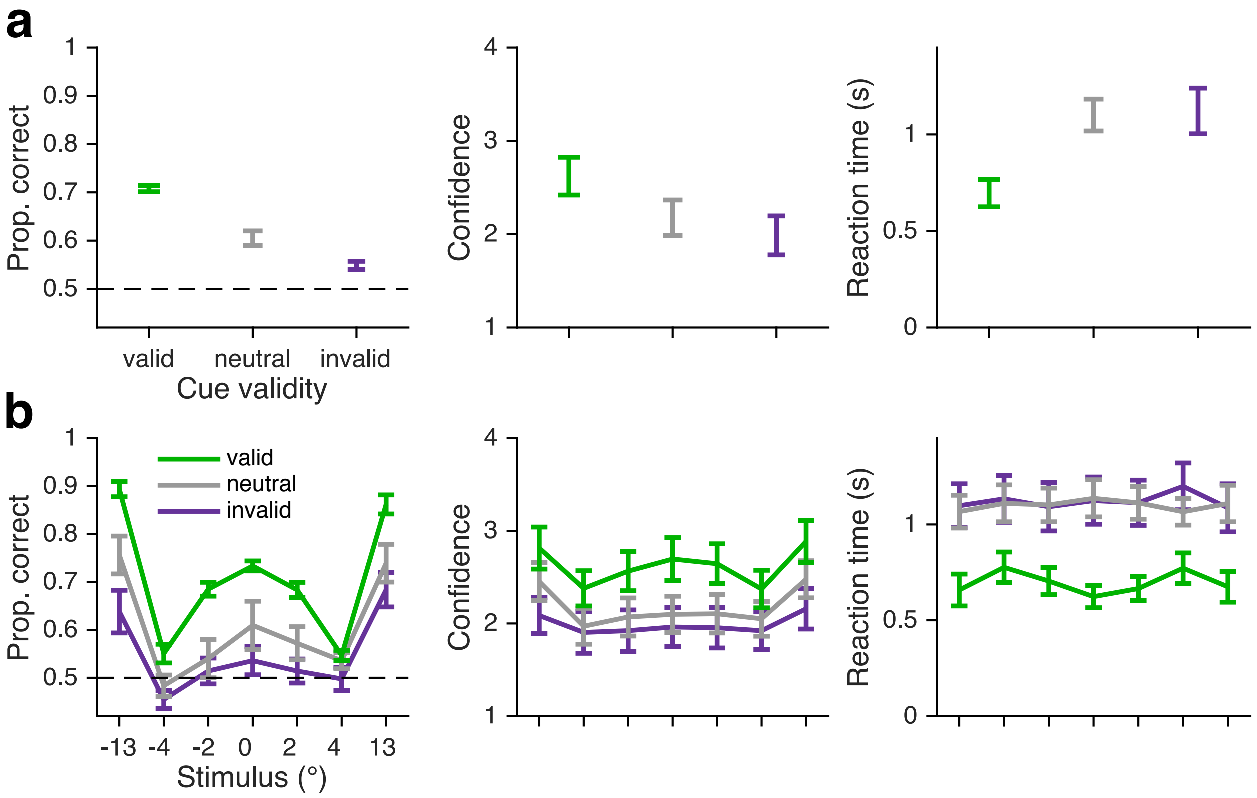

Figure 4.3 Behavioral data. $n = 12$ subjects. Error bars show trial-weighted mean and SEM across subjects. (**a**) Accuracy, confidence ratings, and reaction time as a function of cue validity. Maximum accuracy is ~80% because the stimulus distributions overlap. (**b**) As in **a**, but as a function of cue validity and stimulus orientation.

reflecting the greater difficulty in categorizing a stimulus when its orientation was near a category boundary. Attention increased the sensitivity of category and confidence responses to the stimulus orientation (Figure 4.3b).

4.3.2 Model comparison

To assess whether subjects changed their category and confidence decision boundaries to account for attention-dependent orientation uncertainty, we first fit two of the models described in Chapter 3: Bayesian and Fixed. As described previously, both models assume that, for the stimulus of interest, the observer draws a noisy orientation measurement from a normal distribution centered on the true stimulus value, with s.d. (i.e., uncertainty) dependent on attention. In the Bayesian model, decisions depend on the relative posterior probabilities of the two categories, leading the observer to adjust their decision boundaries in measurement space, based on attention condition (Figure 4.4a,b, Figure 4.6). The Bayesian model maximizes accuracy and produces confidence reports that are a function of the posterior probability of being correct. In the Fixed model, observers use the same decision criteria, regardless of the attention condition (Caetta and Gorea, 2010; Gorea et al., 2005; Gorea and Sagi, 2000, 2001, 2002; Morales et al., 2015; Rahnev et al., 2011, 2012b; Zak et al., 2012) (i.e., they are fixed in measurement space, Figure 4.4a,b). We used Markov Chain Monte Carlo sampling to fit the models to raw, trial-to-trial category and confidence responses from each subject.

Subjects' decisions took attention-dependent uncertainty into account. The Bayesian model captured the data well (Figure 4.4c) and substantially outperformed the Fixed model (Figure 4.4c,d), which had systematic deviations from the data (although the fit depends on the full data set, note deviations near zero tilt and at large tilts in Figure 4.4c). To compare models, we used an approximation of leave-one-out cross-validated log likelihood called PSIS-LOO (henceforth LOO) (Vehtari et al., 2015). Bayesian outperformed Fixed by LOO differences (median and 95% CI of bootstrapped mean[I] differences across subjects) of 102 [45, 167]. This implies that the attentional state is available to the decision process and is incorporated into probabilistic representations used to make the decision.

To determine whether Bayesian computations are necessary to produce the behavioral data, we tested two models with heuristic decision rules, previously described in Chapter 3: Linear and Quadratic. These models approximate the Bayesian boundaries (Figure 4.5a) without

[I] Note that in Chapter 3 we used boostrapped summed differences, which partly explains why the differences might appear smaller in this chapter

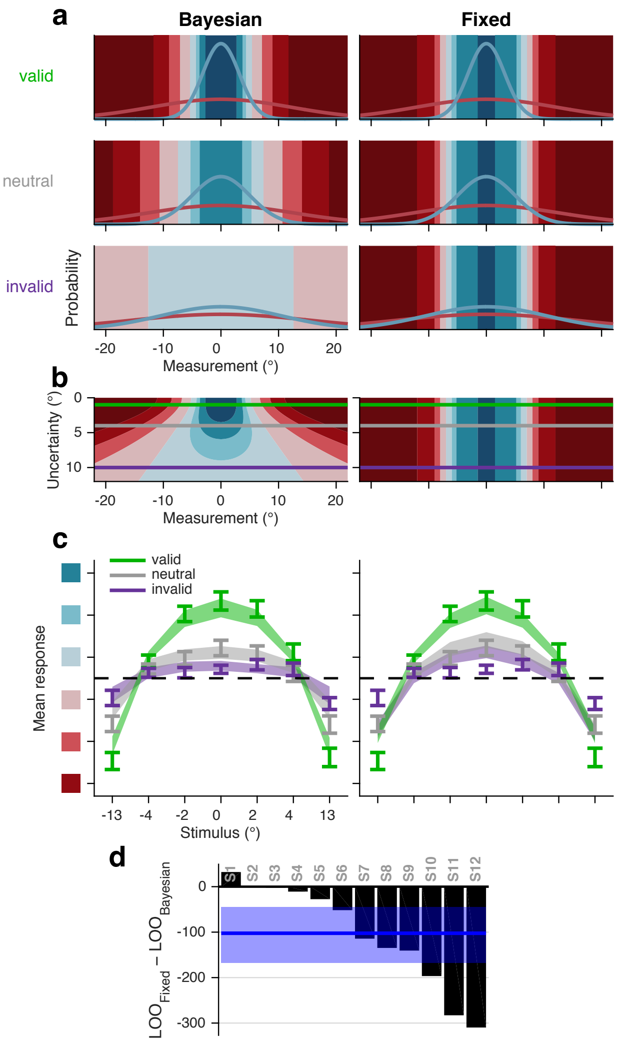

Figure 4.4 Model schematics, fits, and fit comparison. (**a**) Schematic of Bayesian (left) and Fixed (right) models. As attention decreases, uncertainty (the measurement noise s.d.) increases, and orientation measurement likelihoods (blue and red curves) widen (Giordano et al., 2009). In the Bayesian model, choice and confidence boundaries are defined by posterior probability ratios and therefore change as a specific function of uncertainty. In the Fixed model, boundaries do not depend on uncertainty. Colors indicate category and confidence response (color code in Figure 4.1b). (**b**) Decision rules for Bayesian and Fixed models show the mappings from orientation measurement and uncertainty to category and confidence responses. Horizontal lines indicate the uncertainty levels used in **a**; note that the regions intersecting with a horizontal line match the regions in the corresponding plot in **a**. (**c**) Model fits to response as a function of orientation and cue validity. Mean response is an 8-point scale ranging from "high confidence" category 1 to "high confidence" category 2, with colors corresponding to those in Figure 4.1b; only the middle 6 responses are shown. Error bars show mean and SEM across subjects. Shaded regions are mean and SEM of model fits (Methods). Although mean response is shown here, models were fit to raw trial-to-trial data. (**d**) Model comparison. Black bars represent individual subject LOO differences of Bayesian from Fixed. Negative values indicate that Bayesian had a higher (better) LOO score than Fixed. Blue line and shaded region show median and 95% confidence interval of bootstrapped mean differences across subjects.

any computation of the posterior. The Linear and Quadratic models both outperformed the Fixed model (LOO differences of 124 [77, 177] and 129 [65, 198], respectively; Figure 4.5b,c). The best model, quantitatively, was Quadratic, as in Qamar et al. (2013) and Chapter 3

(Table 4.3 shows pairwise LOO comparisons of all models). Decision rules therefore changed with attention without requiring Bayesian computations.

We next asked whether the category decision boundary alone—regardless of confidence—accounts for attention-dependent uncertainty. We were able to answer this question because, unlike in a traditional left vs. right orientation discrimination task, the optimal category decision boundaries in this task depend on orientation uncertainty (Figure 4.4a,b, Figure 4.6) (Qamar et al., 2013). We fit the four models to the category choice data only and again rejected the Fixed model (Figure 4.7a,b; Table 4.4). We also fit the category choice data with a Free model in which the category decision boundaries varied freely and independently for each attention condition. The estimated boundaries differed between valid and invalid trials (Figure 4.7c, Figure 4.8), with a mean difference of 7.5° (s.d. = 7.8°) [two-tailed paired t-test, $t(11) = 3.33$, $p < 10^{-2}$]. Therefore, category criteria, independent of confidence criteria, varied as a function of attention-dependent uncertainty.

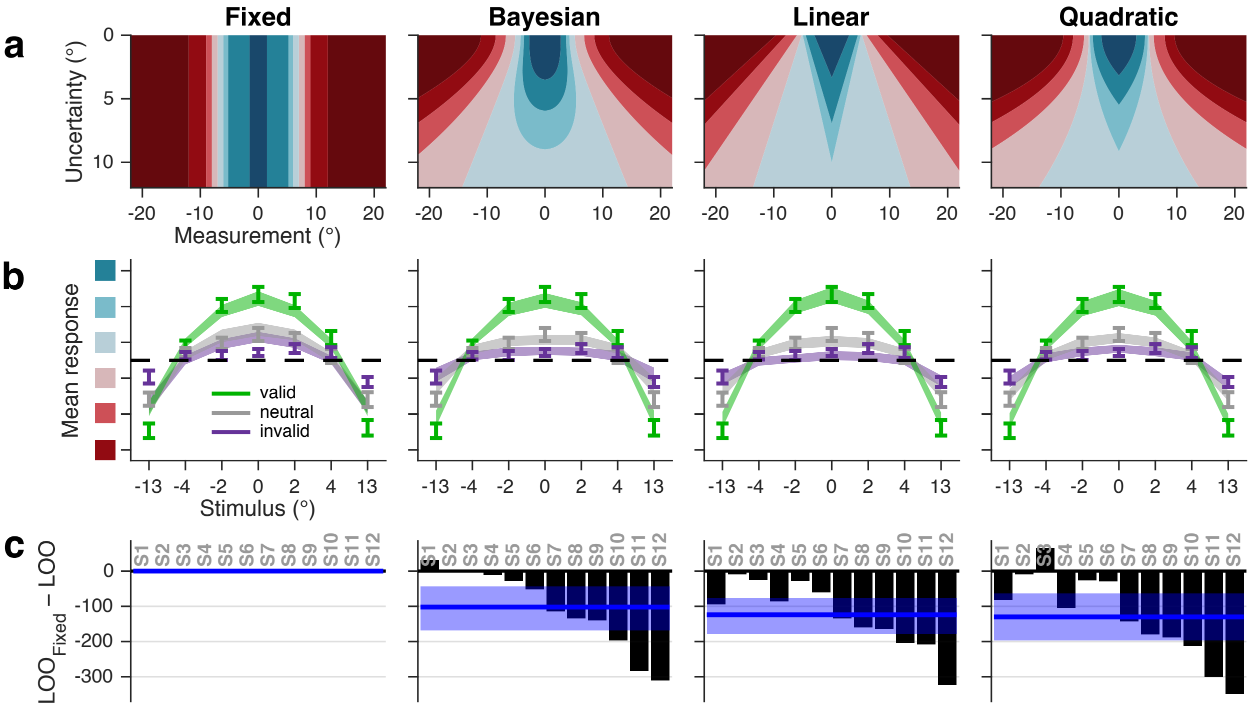

Figure 4.5 Category and confidence models. (**a**) Theoretical relation between orientation uncertainty and category and confidence decision boundaries for all models. (**b**) Mean response as a function of orientation and cue validity, as in Figure 4.4c. (**c**) Model comparison. Black bars represent individual subject LOO score differences of each model from Fixed. Negative values indicate that the corresponding model had a higher (better) LOO score than Fixed. Blue line and shaded region show median and 95% confidence interval of bootstrapped mean LOO differences across subjects.

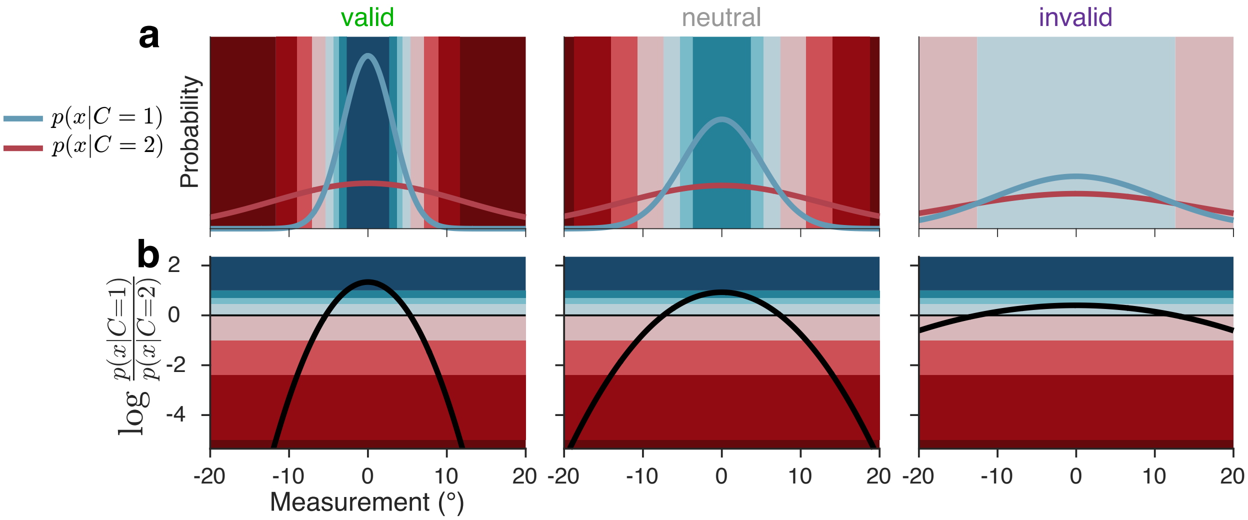

Figure 4.6 The Bayesian mapping from orientation measurement and attention-dependent uncertainty to response. Colors correspond to category and confidence response as in Figure 4.1b. (**a**) Blue and red curves show likelihood functions for the category distributions under example levels of uncertainty. (**b**) The Bayesian model maps measurement and uncertainty onto the decision variable, the log likelihood ratio (black curve). When the relative likelihood of category 1 is high, the decision variable is large and positive; when the relative likelihood of category 2 is high, it is large and negative. Response is determined by comparing the decision variable to boundaries that are fixed in log-likelihood-ratio space, but in measurement space vary as a function of uncertainty.

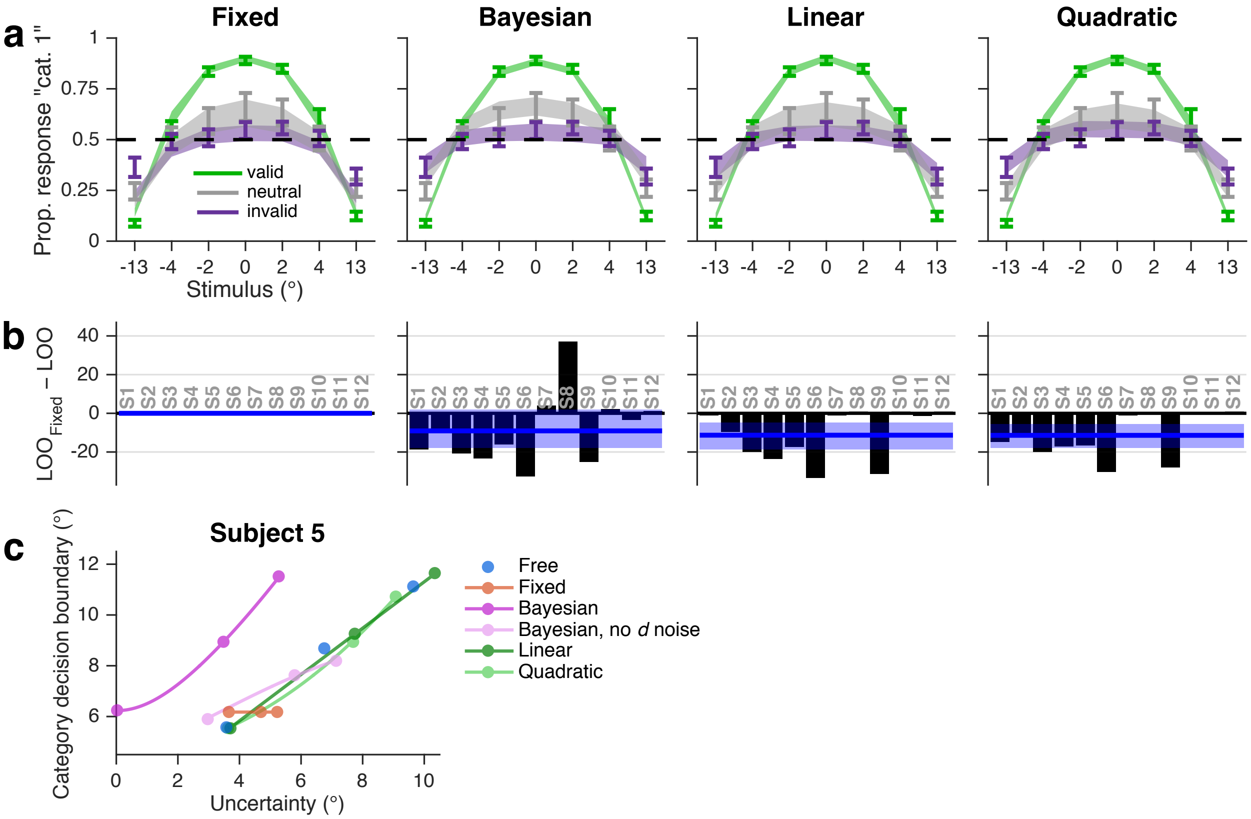

Figure 4.7 Category choice-only models. (**a**) Proportion of category 1 responses as a function of orientation and cue validity. Error bars show mean and SEM across subjects. Shaded regions are mean and SEM of model fits (Methods). (**b**) LOO model comparison, as in Figure 4.5c. (**c**) Mean MCMC orientation uncertainty and category choice boundary parameter estimates for a representative subject. Estimates are plotted as a function of attention condition (valid, neutral, invalid; filled circles), along with their generating functions (curves), for the four main models fit to the category choice data only, plus a Bayesian model with no noise on the decision variable d and a nonparametric model in which choice boundaries are unconstrained (Free; parameter estimates from this model are plotted in gray for all subjects in Figure 4.8). The Bayesian curve is to the left of the other curves, because noise attributed to orientation uncertainty in the other models is partially attributed to decision noise in the Bayesian model; when the decision noise parameter is removed (Bayesian, no d noise), the curve aligns with the others.

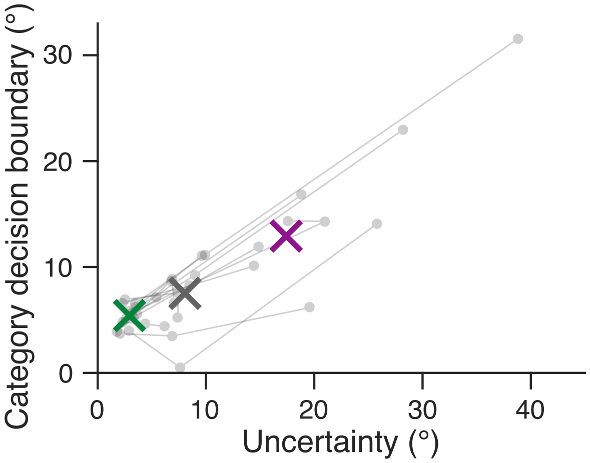

Figure 4.8 Free model analysis. Group mean MCMC parameter estimates (crosses) show systematic changes in the category decision boundary across attention conditions (green = valid, gray = neutral, purple = invalid). The same pattern can be seen for individual subjects: each gray line corresponds to a subject, with connected gray points representing the estimates for valid, neutral, and invalid attention conditions, in that order. Each point represents a pair of parameter estimates: uncertainty and category decision boundary for a specific attention condition.

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

Table 4.3 Cross comparison of all category choice and confidence decision models. Cells indicate medians and 95% CI of bootstrapped mean LOO score differences. A positive median indicates that the model in the corresponding row had a higher score (better fit) than the model in the corresponding column.

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

Table 4.4 Cross comparison of all category choice-only decision models. Conventions as in Table 4.3.

4.3.3 Model comparison metric analysis

We determined that our results were not dependent on our choice of model comparison metric. We computed AIC, BIC, AICc, WAIC (Gelman et al., 2013), and LOO for all models in the 2 model groupings (category choice-plus-confidence and category choice-only), multiplying the non-LOO metrics by $-\frac{1}{2}$ to match the scale of LOO. For AIC, BIC, and AICc, we selected the MCMC sample with the highest log likelihood as our maximum-likelihood parameter estimate. Then we computed Spearman's rank correlation coefficient for every possible pairwise comparison of model comparison metrics for all model and dataset combinations, producing 20 total values (2 model groupings × 10 possible pairwise comparisons of model comparison metrics). All values were greater than 0.998, indicating that, had we used an information criterion instead of LOO, we would not have changed our conclusions. Furthermore, there are no model groupings in which the identities of the lowest- and highest-ranked models are dependent on the choice of metric. The agreement of these metrics strengthens our confidence in our conclusions.

4.3.4 Model recovery

We performed a model recovery analysis (van den Berg et al., 2014) to test our ability to distinguish our choice and confidence models. We generated synthetic datasets from each model, using the same sets of stimuli that were originally randomly generated for each of the 12 subjects. To ensure that the statistics of the generated responses were similar to those of the subjects, we generated responses to these stimuli from 8 of the randomly chosen parameter estimates obtained via MCMC sampling for each subject and model. In total, we generated 384 datasets (4 generating models × 12 subjects × 8 datasets). We then fit all four models to every dataset, using maximum likelihood estimation (MLE) of parameters by an interior-point constrained optimization (MATLAB's *fmincon*), and computed AIC scores from the resulting fits. For reasons of computational tractability, we used AIC instead of LOO as the model comparison metric. Because AIC and LOO scores gave us near-identical model rankings for data from real subjects (Section 4.3.3), we do not believe that the model recovery results are dependent on choice of metric.

We found that the true generating model was the best-fitting model, on average, in all cases (Figure 4.9). Overall, AIC "selected" the correct model (i.e., AIC scores were lowest for the model that generated the data) for 87.5% of the datasets, indicating that our models are distinguishable.

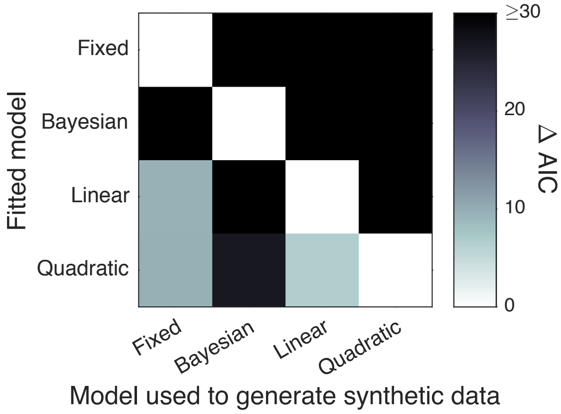

Figure 4.9 Model recovery analysis. Shade represents the difference between the mean AIC score (across synthetic datasets) for each fitted model and for the one with the lowest mean AIC score. White squares indicate the model that had the lowest mean AIC score when fitted to data generated from each model. The fact that all white squares lie on the diagonal indicates that the true generating model was the best-fitting model, on average, in all cases.

4.4 Discussion

In Chapter 3 we found that although subjects use their sense of uncertainty when reporting confidence, they did not do so in a Bayesian way. We wanted to find out if these results held even when sensory uncertainty was due to top-down, rather than bottom-up, stimulus uncertainty. Because we were able to reject the Fixed model here, we conclude that subjects do use their knowledge of top-down uncertainty. However, we were unable to distinguish the Bayesian model from probabilistic non-Bayesian models Linear and Quadratic.

Our rejection of the Fixed model may be surprising in light of the “unified criterion” account of perceptual decision-making (Gorea and Sagi, 2000, 2001). According to this account, when multiple relevant items are simultaneously present, the observer adopts a single, fixed decision boundary (the “unified criterion”) that is used for all items, regardless of stimulus properties or attentional state. Findings considered to support this account (Caetta and Gorea, 2010; Gorea et al., 2005; Gorea and Sagi, 2000, 2001, 2002; Morales et al., 2015; Rahnev et al., 2011, 2012a,b; Zak et al., 2012) imply a rigid, suboptimal mechanism for perceptual decision-making in real-world complex scenes. We consider two possible differences between those studies and ours. First, previous studies (Gorea and Sagi, 2000, 2001) may have been limited in their ability to distinguish changes in criteria from changes in internal signal variability (Kontsevich et al., 2002). Second, the type of perceptual decision may be relevant for an observer’s ability to take uncertainty into account. Our study required subjects to make a categorization decision, whereas those supporting a unified criterion required detection or orthogonal discrimination (Caetta and Gorea, 2010; Gorea et al., 2005; Gorea and Sagi, 2000, 2001, 2002; Morales et al., 2015; Rahnev et al., 2011, 2012b; Zak et al., 2012), which is often used as a proxy for detection (Carrasco et al., 2000; Thomas and Gille, 1979). Additionally, our study asked for confidence rather than visibility (Rahnev et al., 2011); these prompts are known to produce different results (Rausch and Zehetleitner, 2016).

This chapter represents the first study to show that perceptual decisions, including confidence reports, take into account attention-related uncertainty. Only a handful of studies have examined any influence of attention on confidence, and their findings have been mixed. Two studies found that attention increased confidence (Zizlsperger et al., 2012, 2014), but another found no effect (Wilimzig et al., 2008). The latter result has been attributed to response speed pressures (Zizlsperger et al., 2012). Three other studies suggested an inverse relation between attention and confidence: one reported higher confidence for uncued compared to cued error trials (Baldassi et al., 2006), one found higher confidence for stimuli with incongruent compared to congruent flankers (Schoenherr et al., 2010), and a third found that lower fMRI BOLD activation in the dorsal attention network correlated with higher confidence (Rahnev et al., 2012b). Our results, based on an experimental manipulation of spatial attention with no response speed pressure, support a positive relation between spatial attention and confidence and further reveal that it is approximately Bayesian. Attention is typically spread unevenly across multiple objects in a visual scene, so the ability to account for attention likely improves perceptual decisions in natural vision.

The biggest difference between the results presented here and in Chapter 3 is that here, we are unable to distinguish the Bayesian model from the probabilistic non-Bayesian heuristic models Linear and Quadratic. One explanation for this difference is that confidence ratings are more Bayesian under top-down stimulus uncertainty than under bottom-up stimulus uncertainty. However, we think there is a more mundane explanation. Because our models are distinguished by how they take uncertainty into account, model distinguishability increases with the number of uncertainty levels. In Chapter 3, we tested subjects under 6 different uncertainty levels (of contrast or of ellipse elongation). However, here, we tested subjects only under 3 different uncertainty levels (valid, neutral, and invalid attention conditions) because of the difficulty of training subjects to use multiple levels of cue validity. We think that this hurt our ability to distinguish models. Future studies aiming to distinguish these models could obtain more levels of cue validity by using a less coarse cuing mechanism. Another option would be to optimize stimulus distributions and uncertainty levels (via cue validity or contrast) for maximum model distinguishability (Myung and Pitt, 2009).[II]

[II] Here, we optimized contrast to maximize differences in performance levels across attention conditions (Section 4.2.1.3). We were using performance differences as a proxy for uncertainty differences, a prerequisite for distinguishing our models. Optimizing directly for model distinguishability might be more effective.

Chapter 5

Confidence reports from trained neural networks

5.1 Introduction

The outcomes of our model comparisons in the previous chapters do not convince us that confidence ratings are Bayesian. On the contrary, Bayesian models performed markedly worse than heuristic models under bottom-down stimulus uncertainty (Chapter 3) and were indistinguishable from heuristics under top-down uncertainty (Chapter 4). However, one might still conclude, after examining the fits of the Bayesian model, that the behavior is "approximately Bayesian" rather than "non-Bayesian." As written, this is a semantic distinction because it relies on one's definition of "approximate." However, it can be turned into a more meaningful question: Are the differences between human behavior and Bayesian models accounted for by an unknown principle, such as an ecologically relevant objective function that includes both task performance and biological constraints?

Although there are benefits associated with veridical explicit representations of confidence (Bahrami et al., 2012; Bang et al., 2014; Folke et al., 2016), there are also neural constraints that may give rise to non-Bayesian behavior (Bowers and Davis, 2012; Jones and Love, 2011).

Such constraints include the kinds of operations that neurons can perform, the high energy cost of spiking (Attwell and Laughlin, 2001; Lennie, 2003), and the cost of neural wiring length (Chklovskii and Koulakov, 2004; Clune et al., 2013). Perhaps such constraints also restrict the brain's ability to perform fully Bayesian computation. A search for ecologically rational constraints on Bayesian computation benefits from the positive characterization of the deviations from Bayesian computation that we have provided in Chapters 3 and 4, in the form of heuristic models such as Lin and Quad.

Specifically, one possible way to explain confidence ratings under a normative framework is to consider whether the brain might be performing near-optimally given implementational constraints. To explore this possibility, we adopt an approach related to work by Yamins et al. (2014), who trained neural networks on visual categorization tasks, maximizing their performance, without exposing the networks to neural data. They then compared the activation of intermediate layers of the trained networks to the neural activation patterns of midlevel visual areas. They found that the activation of the networks was highly predictive of actual neural response, indicating that the brain may have evolved to maximize performance on similar categorization tasks. Here, we trained simple feedforward neural networks on Task B, as if the networks were naïve human subjects; we did not expose the networks to human behavioral data. We will compare the output of the trained networks to that of the human subjects in Chapter 3 on the basis of model rankings. To anticipate our results, as with the behavior of the human subjects in Chapter 3, we find that the heuristic models, not the Bayesian models, provide the best description of the behavior of the neural networks.

This chapter may open a new research program in which the behavior of neural networks on psychological tasks is systematically compared to the behavior of humans on the basis of model comparison.

5.2 Methods

5.2.1 Architecture

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

As in Orhan and Ma (2017), we trained 3-layer feedforward neural networks to perform Task B (Section 3.2.1). The architecture, described below, is pictured in Figure 5.1. 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

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

where $\tilde{\mathbf{s}}$ 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. Input units were connected, all-to-all, to 200 hidden rectified linear units with responses

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

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

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

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

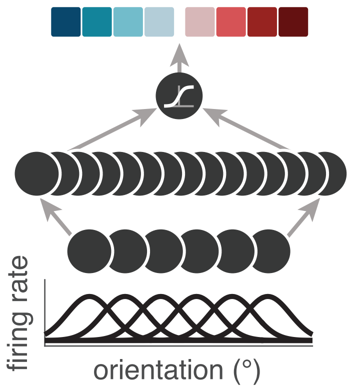

Figure 5.1 Feedforward neural network architecture. Input units were independent Poisson neurons with Gaussian tuning curves that were evenly spaced and of identical width. Input units were connected, all-to-all, to hidden rectified linear units. Hidden units were connected to a sigmoidal output unit. The output was mapped onto a category and confidence response (colors as used in the rest of this dissertation, starting with Figure 3.1a) using eight quantiles.

5.2.2 Training networks and generating datasets

Stimuli s were drawn from the same distributions used for the human experiments in Task B (Chapters 3 and 4).

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 calculated from the fits to the Task A choice data in Chapter 3, experiment 1. To calculate the gains, we used the relationship between gain and sensory uncertainty in populations of independent Poisson neurons, as derived in Ma et al. (2006), supplementary section 2.1. As expected, the performance of the networks roughly matched the performance of the subjects (Figure 5.3c). We used 15 different values for the number of training trials, ranging from 10 to 4.6×10^5, logarithmically spaced (Figure 5.2 only depicts results from the most highly trained networks).

We used standard back-propagation (Rumelhart et al., 1986) 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 $\mathbf{r}_{\text{input}}$ (Beck et al., 2011; Orhan and Ma, 2017).

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

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

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 Chapter 3, 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).

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 (Figure 5.3a). We also found that information loss and performance went down as the number of training trials increased (Figure 5.3b,c). Both information loss and performance appear to reach asymptote around 2×10^4 trials; therefore, it is unlikely that our results would change with more training.

5.2.3 Modeling

We fit the 660 network-generated datasets, obtaining AIC scores for each dataset and model. The models used are described in Chapter 3 (and shown in Figure 3.15 and Table 3.2), 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 our model recovery analyses (Sections 3.3.5 and 4.3.4), 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 computed the expected posterior probability distribution over models at each number of training trials (Figure 5.3d; as described in Section 3.2.4.7).

5.3 Results

We trained biologically plausible feedforward neural networks (Figure 5.1) to perform the task used in Chapters 3 and 4 with online binary category correctness feedback (Orhan and Ma, 2017). After extracting confidence ratings from the networks, we found that confidence behavior is qualitatively similar to that of human subjects (Figure 5.2).

We then fit the network output with the same models that we used to fit subject data. 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 (Figure 5.3d). This transition is consistent with previous results (Orhan and Ma, 2017).

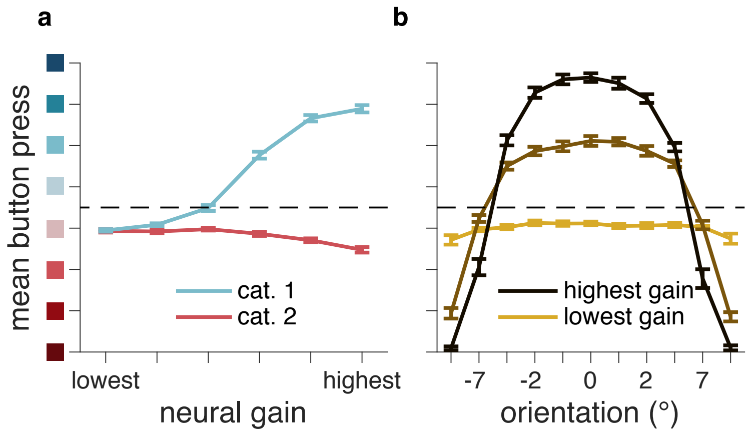

Figure 5.2 Neural network task behavior. (**a**) Mean button press as a function of neural gain and true category. Compare to Figure 3.8d. (**b**) Mean button press as a function of stimulus orientation and neural gain. Compare to Figure 3.8o.

We plotted the summed AIC differences for data generated from the most highly trained networks in Figure 5.4, blue bars. Overall, Lin is the best-fitting model, outperforming $\text{Bayes}_{\text{Weak}}$, the best-fitting Bayesian model, by summed AIC differences of 11662 [10011, 13522]. The overall rankings of all models fit to the trained networks was very similar to that of the human models (Figure 3.15a), with a Spearman's rank correlation coefficient of 0.85. This convergence of neural network and human behavior suggests that neural architecture may impose constraints on the type of behavior that can be produced.

5.3.1 Control

An alternative explanation of this result is that Bayes is too inflexible to fit any behavioral dataset based on neural activity. To rule out this possibility, we used the $\mathbf{r}_{\text{input}}$ from the test set of the 44 most highly trained networks (11 subject-derived sets of gains × 4 datasets). We decoded optimal posterior probabilities from input unit activity $p(C = 1 \mid \mathbf{r}_{\text{input}})$, on a per-trial basis, and mapped these onto button presses using quantiles. 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, fitting these datasets better than Lin by 5739 [3935, 8045] and better than Quad by 800 [479, 1412] (Figure 5.4, black bars). Thus,

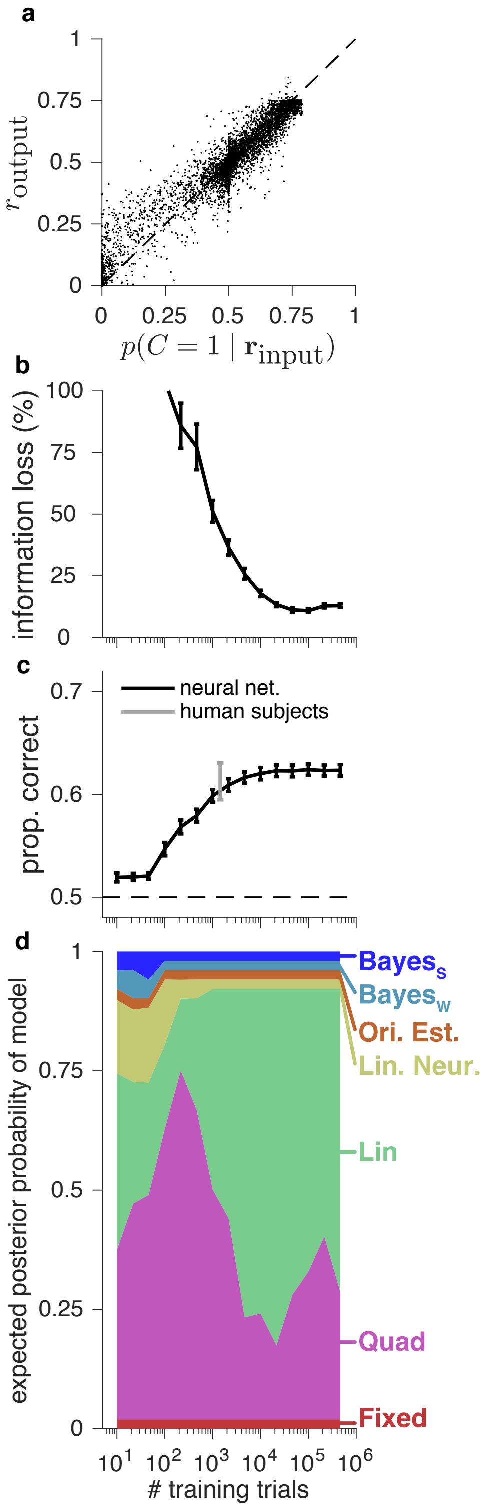

Figure 5.3 Neural network optimality, task performance, and model comparison. (**a**) Posterior probabilities decoded optimally from input unit activity, scattered against network output. Each point is a test trial from a neural network that was trained on the maximum number of training trials. For clarity, 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 probability that a model generated a randomly chosen dataset (Stephan et al., 2009), as a function of the number of training trials.

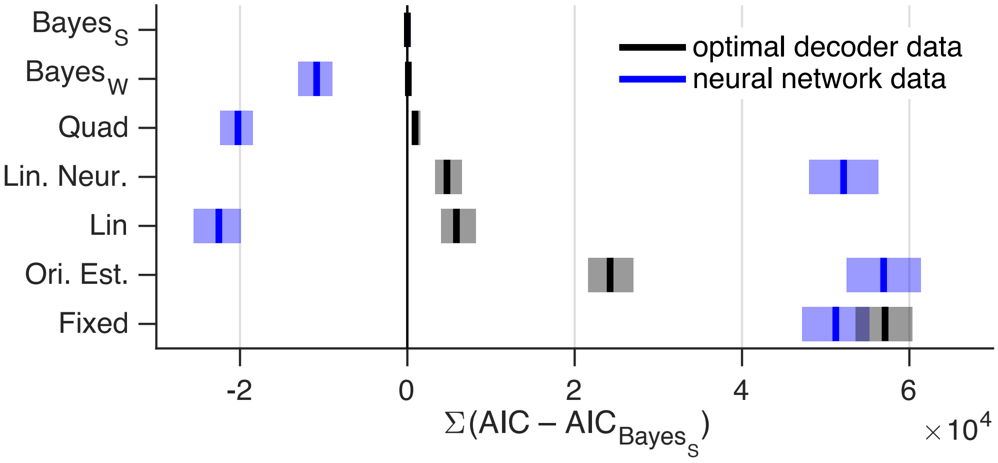

Figure 5.4 Model comparison (as in Figure 3.11) for data generated from an optimal decoder of spikes, and from neural networks trained with 4.6×10^5 training trials. Models are ordered by goodness of fit to optimal decoder data.

the fact that Lin wins is not due to Bayes being generally inflexible, and suggests that the architecture or training procedure of the neural networks constrains the type of behavior that can be produced.

5.4 Discussion

We trained neural networks to perform one of our psychophysical tasks, as if the networks were naïve human subjects. We then fit to the network output using some of the same models that we fit to human data in Chapters 3 and 4. Although the training procedure necessarily differed from that of the humans, we found that the trained networks produced confidence responses that, like the human data in Chapter 3, were best fit by heuristic models. This convergent result suggests that the structure of the neural network—and by extension, the structure of the brain—limits its ability to produce accurate posterior estimates in categorization tasks.

Scientific results are more convincing when paired with a normative explanation. A normative explanation may take the form of an evolutionary just-so story, as in "this result is due to natural selection." Alternatively or complementarily, an explanation may be mechanistic, as in "this result is due to an upstream mechanism." When a Bayesian model gives the best fit to the data, evolutionary just-so stories are commonly used (Bowers and Davis, 2012). But such explanations are not compatible with results showing that an organism appears to have converged on a sub-optimal strategy. In Chapters 3 and 4, we found that Bayesian models are not the best-fitting models to human data, and it therefore may be useful to probe the mechanism to determine what might be causing this result.

Unfortunately, determining the actual neural mechanisms responsible for a behavior is a long and difficult endeavor. This chapter, along with Orhan and Ma (2017), is an early test of a possible way to probe the neural mechanisms of behavior, in silico, without laborious experimentation. One possible future research program would be to, for a given psychophysical task, thoroughly explore the space of behavioral models and neural network architectures and training procedures. As these components are varied, one could measure the similarity of trained network behavior to that of humans on the basis of model rankings, perhaps by using Spearman's rank correlation coefficient. This would be analogous to how Yamins et al. (2014) vary neural network architecture, measuring similarity to real neural activation by using percent explained variance in inferotemporal cortex.

After determining a network architecture that produces human-like behavior, the next challenge would be to probe the internal workings of the trained networks. A parallel challenge has been faced by machine learning and neural network researchers in general: some categories of models perform better than others, but at a mechanistic level it is not always clear why. In image classification, recent methods have been used for visualizing the representations and functions performed by intermediate layers in deep networks (Dosovitskiy and Brox, 2015; Mahendran and Vedaldi, 2015; Simonyan et al., 2013). Similar methods could be developed for understanding why a trained network produces some behavior. One could also add or remove various components of the network to see which components are most associated with human-like deviations from the Bayesian model.

We intend this chapter to be a small advance that points towards one possible way of analyzing neural networks from a different behavioral angle than has been done in the past. This approach could aid in the formation of hypotheses about mechanistic explanations for behavior, which could then be tested in vivo.

Chapter 6

Conclusion

In this dissertation, we have studied explicit human confidence ratings in perceptual categorization. We have used formal model comparison to distinguish a large set of computational models of confidence, with a particular focus on testing whether confidence ratings are Bayesian.

Previous work has proposed that confidence should be defined as Bayesian (Kepecs and Mainen, 2012; Meyniel et al., 2015; Pouget et al., 2016). However, the notion that confidence is Bayesian is not an established fact but a hypothesis, which we call the Bayesian Confidence Hypothesis (BCH). The work presented here represents the most comprehensive test of the BCH to date.

We opened this work by analyzing a proposed approach for studying confidence that involves qualitative "signatures" of confidence (Chapter 2). We concluded that this approach is unable to determine whether confidence ratings are Bayesian, and so we instead use quantitative model comparison. In Chapter 3, we used a set of binary categorization tasks in which we induced sensory uncertainty by manipulating stimulus factors such as contrast. Chapter 4 was an extension of Chapter 3 in which we induced sensory uncertainty by manipulating subjects' attention. We concluded from Chapters 3 and 4 that there is mixed evidence for the BCH. Qualitatively, it appears that people take their sensory uncertainty into account in an approximately Bayesian way. But quantitatively, heuristic models perform as well as, or better than, the Bayesian models we tested. In Chapter 5, we trained simple neural networks in an attempt to understand more about the origin of heuristic computations.

6.1 Potential caveats

6.1.1 Confidence behavior

Our results in Chapters 3 and 4 may be affected by two design choices we made in order to obtain a naïve confidence report uninfluenced by reward or experimenter instruction. The first design choice was to not incentivize confidence reports. As with other work (Aitchison et al., 2015; Navajas et al., 2017; Sanders et al., 2016), this means that there was arguably no reason for subjects to report Bayesian confidence. Future research could extend the experiments conducted here, and investigate whether Bayesian models outperform heuristics when Bayesian confidence reports is incentivized. Other techniques for measuring confidence, such as post-decision wagering (Persaud et al., 2007), may be useful here. But, among other issues (Grimaldi et al., 2015), using techniques that utilize reward may train observers to use a particular model of confidence. In other words, a researcher who rewards some confidence reports might be asking "can I train an observer's confidence to be Bayesian?" instead of "is confidence Bayesian?" We do not know of a way out of this paradox. Future researchers should design a technique to collect confidence reports in which subjects have a reason to provide Bayesian confidence reports, but are not incidentally trained to do so. The second design choice was to not explicitly define "confidence" for the subject. Recent research has indicated that the language used by the experimenter may influence confidence reports (Rausch and Zehetleitner, 2016). Future researchers should investigate whether results depend on the subject prompt. For instance, instead of asking merely for "confidence," we could have asked subjects to report "confidence that the choice is correct." And in a ternary (or more) categorization task, one could also ask subjects to report “confidence that the choice is better than the next best choice,” which would be a different quantity.

In this study, we only considered explicit confidence ratings, which differ from the implicit confidence that can be gathered from nonhuman animals (Kepecs and Mainen, 2012) (e.g., by measuring how frequently they decline to make a difficult choice (Kiani and Shadlen, 2009), or how long they will wait for a reward (Kepecs et al., 2008)). It is possible that implicit confidence might be more Bayesian (Chen et al., 2014). At a minimum, testing this possibility would require an experiment using implicit confidence that could distinguish the models presented here, which has not been done.

6.1.2 Modeling

In Chapters 3 to 5, the best-fitting models tend to have more parameters, raising the question of whether our winning models are merely overfitting the data. We have tried to avoid this issue as much as as possible by using a wide variety of model comparison metrics, including those that approximate leave-one-out cross-validation (Section 3.2.4.7), but this may not have been enough. Because real behavioral data is so complex, a highly-parameterized model may fit better, even when the model is properly penalized for complexity, because it really is closer to the complex model used by the organism. So even with proper model comparison techniques, it may not be surprising that models with more parameters win out. This also poses a problem for the conclusions that can be drawn from the neural network analysis in Chapter 5. Perhaps the strong performance of the heuristic models in all three chapters tells us more about the models themselves than about the behavior from the humans or the networks.

6.2 Topics not addressed

This dissertation concerns itself only with binary categorization tasks. Future confidence research should focus on more naturalistic tasks, including categorization tasks with more than two options. In real life, binary decisions may be less common than decisions that involve choosing from many discrete categories, or along a continuous axis. Some recent confidence work has taken a step in this direction by using a ternary categorization task (Li and Ma, 2017).

This dissertation also does not address the question of confidence calibration (Baranski and Petrusic, 1994; Brier, 1950). An old question in confidence is whether humans can report well-calibrated, veridical confidence, i.e., confidence reports that are not only a function of, but equal to the true probability of being correct. As described above, we did not explicitly define "confidence" for the subject. However, if we had defined "confidence" as "probability that a choice is correct," and assigned probability ranges to each button (asking subjects, for example, to press the "high confidence" button when their confidence was above 90%), we could have tested how well-calibrated subjects are. If you tell me that you have read this sentence, I will buy you a beer. In our model comparison framework, we could have tested subject calibration by then fitting a model even stronger than $\text{Bayes}_{\text{Ultrastrong}}$ (Chapter 3), with confidence boundary parameters (Section 3.2.4.3) fixed corresponding to the probability ranges described to the subjects. Given the poor performance of $\text{Bayes}_{\text{Ultrastrong}}$ relative to models with fewer constraints, we think it is unlikely that an even more constrained model would do well; however, specific subject instructions may have a strong effect on the results.

Another topic that we do not cover is drift-diffusion models, which have become popular in the confidence literature in recent years due to their ability to explain interactions between choice, confidence, and reaction time, within a single framework (Kiani et al., 2014; Kiani and

Shadlen, 2009; Pleskac and Busemeyer, 2010). Such models are typically used to describe behavior in response to stimuli that vary in duration, usually because the subject is able to terminate stimulus presentation. However, in these sorts of tasks, optimality is difficult to characterize (Drugowitsch et al., 2014a), making it harder to test the BCH, our primary goal. In our data, reaction times are more or less constant; they do not vary as a function of confidence and category response, accuracy, or stimulus difficulty (Figure 3.7). Therefore, a drift-diffusion model is unlikely to do a better job of explaining our results than the static models used here.

6.3 Interpretation

One could take two different views of our heuristic model results in Chapters 3 and 4. The first view is that the heuristics should be taken seriously as principled models (Gigerenzer et al., 2011); here, the challenge is to demonstrate that they describe behavior in a variety of tasks and can be motivated based on underlying principles. The second view is that these are descriptive models simply meant to demonstrate that a simple model can provide a good fit to the data; here, the heuristics are benchmarks for more principled models, and the challenge is to find a principled model that fits the data as well as the heuristics. We lean towards the second view and interpret our results as demonstrating that the BCH may not be the best description of human confidence reports.

This conclusion might be unsatisfying to a reader in search of positive evidence of some principled model. Although we do not think that such evidence is necessary to draw conclusions about confidence reports, we of course see the appeal. To increase the odds of finding positive support of a principled model, an experiment must be specifically designed to distinguish multiple principled models; it is not enough to just test one and show a reasonable fit. Again, it may help here to use tasks with more than two categories. A ternary categorization task, unlike a binary categorization task, is able to distinguish the following three models: the posterior probability of the chosen category, the entropy of the posterior distribution over categories, and the difference in probability between the most and second-most probable categories; Li and Ma (2017) find evidence for the last model, arguably the least principled model of the three.

What do our findings tell us about the neural basis of confidence? Previous studies have found that neural activity in some brain areas (e.g., human medial temporal lobe (Rutishauser et al., 2015) and prefrontal cortex (Fleming et al., 2012), monkey lateral intraparietal cortex (Kiani and Shadlen, 2009) and pulvinar (Komura et al., 2013), rodent orbitofrontal cortex (Kepecs et al., 2008)) is associated with behavioral indicators of confidence, and/or with the distance of a stimulus to a decision boundary. However, such studies mostly used stimuli that vary along a single dimension (e.g., net retinal dot motion energy, mixture of two odors). Because measurement is indistinguishable from the probability of being correct in these classes of tasks, neural activity associated with confidence may represent either the measurement or the probability of being correct (Aitchison et al., 2015). In addition to the recommendation of Aitchison et al. (2015) to distinguish between these possibilities by varying stimuli along two dimensions, we recommend fitting both Bayesian and non-Bayesian probabilistic models to behavior. In view of the relatively poor performance of the Bayesian models in Chapter 3, the proposal (Pouget et al., 2016) to correlate behavior and neural activity with predictions of the Bayesian confidence model should be viewed with skepticism.

There are many neural constraints that may give rise to non-Bayesian behavior, which we have described in Section 5.1. These constraints may be responsible for behaviors in which humans seem to adopt a non-Bayesian solution that “satisfices” (Bowers and Davis, 2012; Jones and Love, 2011; Simon, 1956). Our results show that confidence reports may be one additional behavior in which humans “satisfice.”

We close with some general thoughts on confidence research. As is the case in many scientific fields, the literature contains a substantial number of conflicting results. However, our ability to resolve these conflicts seems to be hampered by several factors. First, behavior seems to be affected by relatively minor changes in experimental paradigm, such as confidence report type (Aitchison and Latham, 2014; Rausch and Zehetleitner, 2016, although not in Chapter 3), which complicates our ability to compare results across studies. We think that differences across confidence report type is unlikely to have real-world significance, and may not be worthy of much future study. Second, there are different classes of models which may not be equally applicable to all experiments. For instance, as described in Section 6.2, drift-diffusion models may not be useful for explaining confidence behaviors in all situations. In our opinion, there has been little effort put into unifying disparate experimental and theoretical work. And finally, the interpretation of confidence behavior is made difficult by the three-way tension between eliciting "natural" confidence reports, motivating subjects to respond accurately, and avoiding training them to respond according to a particular model (Section 6.1.1); this tension is even greater in nonhuman animal research. Given all this, we are not very optimistic about the hopes for a better understanding of confidence reports, unless the field undergoes some major unforeseen transformation. In conclusion, we take a different stance than some of our colleagues (Drugowitsch, 2016); we are not at all confident in our understanding of the nature of confidence.

References

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

Acerbi, L., Vijayakumar, S., and Wolpert, D. M. (2014). On the origins of suboptimality in human probabilistic inference. *PLoS Computational Biology*, 10(6):e1003661.

Acerbi, L., Wolpert, D. M., and Vijayakumar, S. (2012). Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Computational Biology*, 8(11):e1002771.

Aitchison, L., Bang, D., Bahrami, B., and Latham, P. E. (2015). Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Computational Biology*, 11(10):e1004519.

Aitchison, L. and Latham, P. E. (2014). Bayesian synaptic plasticity makes predictions about plasticity experiments in vivo. *arXiv*.

Anton-Erxleben, K. and Carrasco, M. (2013). Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence. *Nature Reviews Neuroscience*, 14(3):188–200.

Ashby, F. G. and Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1):33–53.

Attwell, D. and Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism*, 21(10):1133–1145.

Augsburger, J. J., Corre a, Z. l. M., Trichopoulos, N., and Shaikh, A. (2008). Size overlap between benign melanocytic choroidal nevi and choroidal malignant melanomas. *Investigative Ophthalmology and Visual Science*, 49(7):2823–6.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., and Frith, C. (2012). What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1350–1365.

Bahrami, B., Olsen, K., Latham, P. E., and Roepstorff, A. (2010). Optimally interacting minds. *Science*, 329(5995):1081–1085.

Baldassi, S., Megna, N., and Burr, D. C. (2006). Visual clutter causes high-magnitude errors. *PloS Biology*, 4(3):e56.

Bang, D., Fusaroli, R., Tylén, K., Olsen, K., Latham, P. E., Lau, J. Y. F., Roepstorff, A., Rees, G., Frith, C. D., and Bahrami, B. (2014). Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and Cognition*, 26:13–23.

Baranski, J. V. and Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception and Psychophysics*, 55(4):412–428.

Beck, J. M., Latham, P. E., and Pouget, A. (2011). Marginalization in neural circuits with divisive normalization. *Journal of Neuroscience*, 31(43):15310–15319.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., and Pouget, A. (2012). Not noisy, just wrong: The role of suboptimal inference in behavioral variability. *Neuron*, 74(1):30–39.

Bowers, J. S. and Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3):389–414.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4):433–436.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Britten, K. H., Shadlen, M. N., Newsome, W. T., and Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, 12(12):4745–4765.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, 109(2):204–223.

Caetta, F. and Gorea, A. (2010). Upshifted decision criteria in attentional blink and repetition blindness. *Visual Cognition*, 18(3):413–433.

Cameron, E. L., Tai, J. C., and Carrasco, M. (2002). Covert attention affects the psychometric function of contrast sensitivity. *Vision Research*, 42(8):949–967.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, 51(13):1484–1525.

Carrasco, M., Penpeci-Talgar, C., and Eckstein, M. (2000). Spatial covert attention increases contrast sensitivity across the CSF: Support for signal enhancement. *Vision Research*, 40(10-12):1203–1215.

Chen, S. Y., Ross, B. H., and Murphy, G. L. (2014). Implicit and explicit processes in category-based induction: Is induction best when we don't think? *Journal of Experimental Psychology: General*, 143(1):227–246.

Chklovskii, D. B. and Koulakov, A. A. (2004). Maps in the brain: What can we learn from them? *Annual Review of Neuroscience*, 27:369–392.

Clune, J., Mouret, J. B., and Lipson, H. (2013). The evolutionary origins of modularity. *Proceedings of the Royal Society B: Biological Sciences*, 280:20122863.

Denison, R. N. (2017). Precision, not confidence, describes the uncertainty of perceptual experience: Comment on John Morrison's "Perceptual Confidence". *Analytic Philosophy*, 58(1):58–70.

DiMattina, C. (2016). Comparing models of contrast gain using psychophysical experiments. *Journal of Vision*, 16(9):1–18.

Dosovitskiy, A. and Brox, T. (2015). Inverting visual representations with convolutional networks. *arXiv*.

Drugowitsch, J. (2016). Becoming confidence in the statistical nature of human confidence judgments. *Neuron*, 90(3):425–427.

Drugowitsch, J., DeAngelis, G. C., Klier, E. M., Angelaki, D. E., and Pouget, A. (2014a). Optimal multisensory decision-making in a reaction-time task. *eLife*, 3:e03005.

Drugowitsch, J., Moreno-Bote, R., and Pouget, A. (2014b). Relation between belief and performance in perceptual decision making. *PLoS ONE*, 9(5):e96511.

Fetsch, C. R., Kiani, R., Newsome, W. T., and Shadlen, M. N. (2014). Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron*, 83(4):797–804.

Fleming, S. M. and Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, 124(1):91–114.

Fleming, S. M. and Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1338–1349.

Fleming, S. M., Huijgen, J., and Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *Journal of Neuroscience*, 32(18):6117–6125.

Fleming, S. M. and Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8:443.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., and Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, 329(5998):1541–1543.

Folke, T., Jacobsen, C., Fleming, S. M., and De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, 1(2).

Foote, A. L. and Crystal, J. D. (2007). Metacognition in the rat. *Current Biology*, 17(6):551–555.

Frith, C. D. and Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*, 63(1):287–313.

Gelman, A., Hwang, J., and Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

Gigerenzer, G., Hertwig, R., and Pachur, T. (2011). *Heuristics: The foundations of adaptive behavior*. Oxford: Oxford University Press.

Giordano, A. M., McElree, B., and Carrasco, M. (2009). On the automaticity and flexibility of covert attention: A speed-accuracy trade-off analysis. *Journal of Vision*, 9(3):30.1–10.

Girshick, A. R., Landy, M. S., and Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, 14(7):926–932.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gorea, A., Caetta, F., and Sagi, D. (2005). Criteria interactions across visual attributes. *Vision Research*, 45(19):2523–2532.

Gorea, A. and Sagi, D. (2000). Failure to handle more than one internal representation in visual detection tasks. *Proceedings of the National Academy of Sciences*, 97(22):12380–12384.

Gorea, A. and Sagi, D. (2001). Disentangling signal from noise in visual contrast discrimination. *Nature Neuroscience*, 4(11):1146–1150.

Gorea, A. and Sagi, D. (2002). Natural extinction: A criterion shift phenomenon. *Visual Cognition*, 9(8):913–936.

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

Grimaldi, P., Lau, H., and Basso, M. A. (2015). There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making. *Neuroscience and Biobehavioral Reviews*, 55:88–97.

Hampton, R. R. (2009). Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms? *Comparative Cognition & Behavior Reviews*, 4:17–28.

Hangya, B., Sanders, J. I., and Kepecs, A. (2016). A mathematical framework for statistical decision confidence. *Neural Computation*, 28(9):1840–1858.

Healy, A. F. and Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory*, 7(5):344–354.

Insabato, A., Pannunzi, M., and Deco, G. (2016). Neural correlates of metacognition: A critical perspective on current tasks. *Neuroscience and Biobehavioral Reviews*, 71:167–175.

Jones, M. and Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34(4):169–188.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kepecs, A. and Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1322–1337.

Kepecs, A., Uchida, N., Zariwala, H. A., and Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455(7210):227–231.

Keshvari, S., van den Berg, R., and Ma, W. J. (2012). Probabilistic computation in human perception under variability in encoding precision. *PLoS ONE*, 7(6):e40216.

Kiani, R., Corthell, L., and Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, 84(6):1329–1342.

Kiani, R. and Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, 324(5928):759–764.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., and Murray, R. (2007). What's new in Psychtoolbox-3? *Perception*, 36(14):1–16.

Knill, D. C. and Richards, W. (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T., and Miyamoto, A. (2013). Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature Neuroscience*, 16(6):749–755.

Kontsevich, L. L., Chen, C.-C., Verghese, P., and Tyler, C. W. (2002). The unique criterion constraint: A false alarm? *Nature Neuroscience*, 5(8):707.

Kontsevich, L. L. and Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, 39(16):2729–2737.

Körding, K. (2007). Decision theory: What "should" the nervous system do? *Science*, 318(5850):606–610.

Körding, K. P. and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247.

Koriat, A. (2012). When are two heads better than one and why? *Science*, 336(6079):360–362.

Lak, A., Costa, G. M., Romberg, E., Koulakov, A. A., Mainen, Z. F., and Kepecs, A. (2014). Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron*, 84(1):1–12.

Lee, W. and Janke, M. (1964). Categorizing externally distributed stimulus samples for three continua. *Journal of Experimental Psychology*, 68(1):376–382.

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, 13(6):493–497.

Li, H.-H. and Ma, W. J. (2017). Computation of human confidence reports in decision-making with multiple alternatives. In *2017 Cognitive Computational Neuroscience Meeting*, New York.

Liu, Z., Knill, D. C., and Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research*, 35(4):549–568.

Lu, Z. L. and Dosher, B. A. (1998). External noise distinguishes attention mechanisms. *Vision Research*, 38(9):1183–1198.

Ma, W. J. (2010). Signal detection theory, uncertainty, and Poisson-like population codes. *Vision Research*, 50(22):2308–2319.

Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends in Cognitive Sciences*, 16(10):511–518.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438.

Ma, W. J. and Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, 37(1):205–220.

Maddox, W. T. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior*, 78(3):567–595.

Mahendran, A. and Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5188–5196. IEEE.

Maloney, L. T. and Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience*, 26(1):147–155.

Maniscalco, B. and Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1):422–430.

Massoni, S., Gajdos, T., and Vergnaud, J.-C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*, 5(325):1455.

Meyniel, F., Sigman, M., and Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron*, 88(1):78–92.

Morales, J., Solovey, G., Maniscalco, B., Rahnev, D., de Lange, F. P., and Lau, H. (2015). Low attention impairs optimal incorporation of prior knowledge in perceptual decisions. *Attention, Perception, and Psychophysics*, 77(6):2021–2036.

Moran, R., Teodorescu, A. R., and Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, 78:99–147.

Morrison, J. (2016). Perceptual confidence. *Analytic Philosophy*, 57(1):15–48.

Morrison, J. (2017). Perceptual confidence and categorization. *Analytic Philosophy*, 58(1):71–85.

Myung, J. I. and Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, 116(3):499–518.

Naka, K. I. and Rushton, W. A. (1966). S-potentials from luminosity units in the retina of fish (Cyprinidae). *Journal of Physiology*, 185(3):587–599.

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

Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E., and Bahrami, B. (2017). The idiosyncratic nature of confidence. *Nature Human Behaviour*, 11(11):1–12.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, 31(3):705–767.

Newsome, W. T., Britten, K. H., and Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature*, 341(6237):52–54.

Norton, E. H., Fleming, S. M., Daw, N. D., and Landy, M. S. (2017). Suboptimal criterion learning in static and dynamic environments. *PLoS Computational Biology*, 13(1):e1005304.

Orhan, A. E. and Jacobs, R. A. (2014). Are performance limitations in visual short-term memory tasks due to capacity limitations of model mismatch? *arXiv*.

Orhan, A. E. and Ma, W. J. (2017). Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. *Nature Communications*, 8(1):138.

Palmer, E. C., David, A. S., and Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Consciousness and Cognition*, 28(C):151–160.

Palminteri, S., Wyart, V., and Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, 21(6):425–433.

Peirce, C. S. and Jastrow, J. (1884). On small differences in sensation. *Memoirs of the National Academy of Sciences*, 3:73–83.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10(4):437–442.

Persaud, N., McLeod, P., and Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, 10(2):257–261.

Pleskac, T. J. and Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, 117(3):864–901.

Pouget, A., Drugowitsch, J., and Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, 19(3):366–374.

Prins, N. (2012). The psychometric function: The lapse rate revisited. *Journal of Vision*, 12(6):25–25.

Purcell, B. A. and Kiani, R. (2016). Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proceedings of the National Academy of Sciences*, 113(31):E4531–E4540.

Qamar, A. T., Cotton, R. J., George, R. G., Beck, J. M., Prezhdo, E., Laudano, A., Tolias, A. S., and Ma, W. J. (2013). Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proceedings of the National Academy of Sciences*, 110(50):20332–20337.

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., de Lange, F. P., and Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, 14(12):1513–1515.

Rahnev, D. A., Bahdo, L., de Lange, F. P., and Lau, H. (2012a). Prestimulus hemodynamic activity in dorsal attention network is negatively associated with decision confidence in visual perception. *Journal of Neurophysiology*, 108(5):1529–1536.

Rahnev, D. A., Maniscalco, B., Luber, B., Lau, H., and Lisanby, S. H. (2012b). Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *Journal of Neurophysiology*, 107(6):1556–1563.

Rausch, M. and Zehetleitner, M. (2016). Visibility is not equivalent to confidence in a low contrast orientation discrimination task. *Frontiers in Psychology*, 7:591.

Reynolds, J. H. and Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, 27:611–647.

Rigoux, L., Stephan, K. E., Friston, K. J., and Daunizeau, J. (2014). Bayesian model selection for group studies — Revisited. *NeuroImage*, 84:971–985.

Ross, L. A., Dodson, J. E., Edwards, J. D., Ackerman, M. L., and Ball, K. (2012). Self-rated driving and driving safety in older adults. *Accident Analysis and Prevention*, 48:523–527.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Rutishauser, U., Ye, S., Koroma, M., Tudusciuc, O., Ross, I. B., Chung, J. M., and Mamelak, A. N. (2015). Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nature Neuroscience*, 18(7):1041–1050.

Sanborn, A. N., Griffiths, T. L., and Shiffrin, R. M. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, 60(2):63–106.

Sanders, J. I., Hangya, B., and Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron*, 90(3):499–506.

Schoenherr, J. R., Leth-Steensen, C., and Petrusic, W. M. (2010). Selective attention and subjective confidence calibration. *Attention, Perception, and Psychophysics*, 72(2):353–368.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv*.

Smith, J. D., Beran, M. J., Couchman, J. J., and Coutinho, M. V. C. (2008). The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin & Review*, 15(4):679–691.

Smith, J. D., Schull, J., Strote, J., McGee, K., Egnor, R., and Erb, L. (1995). The uncertain response in the bottlenosed dolphin (Tursiops truncatus). *Journal of Experimental Psychology: General*, 124(4):391–408.

Smith, J. D., Shields, W. E., Schull, J., and Washburn, D. A. (1997). The uncertain response in humans and animals. *Cognition*, 62(1):75–97.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, 46(4):1004–1017.

Sutton, J. E. and Shettleworth, S. J. (2008). Memory without awareness: Pigeons do not show metamemory in delayed matching to sample. *Journal of Experimental Psychology: Animal Behavior Processes*, 34(2):266–282.

Thomas, J. P. and Gille, J. (1979). Bandwidths of orientation channels in human vision. *Journal of the Optical Society of America*, 69(5):652–660.

Thomas, J. P. and McFadyen, R. G. (1995). The confidence heuristic: A game-theoretic analysis. *Journal of Economic Psychology*, 16(1):97–113.

van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., and Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*, 5:e12192.

van den Berg, R., Awh, E., and Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, 121(1):124–149.

Vehtari, A., Gelman, A., and Gabry, J. (2015). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *arXiv*.

Vickers, D. D. (1979). *Decision processes in visual perception.* New York: Academic Press.

Wilimzig, C., Tsuchiya, N., Fahle, M., Einhäuser, W., and Koch, C. (2008). Spatial attention increases performance but not subjective confidence in a discrimination task. *Journal of Vision*, 8(5):7.1–10.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.

Zak, I., Katkov, M., Gorea, A., and Sagi, D. (2012). Decision criteria in dual discrimination tasks estimated using external-noise methods. *Attention, Perception, and Psychophysics*, 74(5):1042–1055.

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

Zizlsperger, L., Sauvigny, T., Händel, B., and Haarmeier, T. (2014). Cortical representations of confidence in a visual perceptual decision. *Nature Communications*, 5:3940.