bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A mathematical framework for statistical decision confidence

Balázs Hangya[1, 2], **Joshua I. Sanders**[1], **Adam Kepecs**[1, *]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 11724, United States.

[2]Laboratory of Cerebral Cortex Research, Institute of Experimental Medicine, Hungarian Academy of Sciences, Budapest, H-1083, Hungary.

[*]kepecs@cshl.edu.

Keywords: Bayesian posterior probability, choice accuracy, evidence discriminability

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Decision confidence is a forecast about the probability that a decision will be correct. For human decision makers, confidence is a deeply subjective sense that can be difficult to study due to its inherently introspective nature. However, confidence can be framed as an objective mathematical quantity – the Bayesian posterior probability, providing a formal definition of statistical decision confidence. Here we use this definition as a starting point to develop a normative statistical framework for decision confidence. We analytically prove interrelations between statistical decision confidence and other observable decision measures. Among these is a counterintuitive property of confidence – that the lowest average confidence occurs when classifiers err in the presence of the strongest evidence. These results lay the foundations for a mathematically rigorous treatment of decision confidence that can lead to a common framework for understanding confidence across different research domains, from human behavior to neural representations.

1 Introduction

Humans and other animals are often faced with decisions with uncertain consequences. In order to successfully cope with these situations, it is critical to be able to compute the correct degree of confidence in a decision. Because confidence is thought to arise from the monitoring of mental content, it is sometimes framed as a form of metacognition associated with other subjective human qualities, such as introspection, awareness

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and even self-reflective consciousness (Charles et al., 2013; Flavell, 1979; Fleming and Lau, 2014; Kunimoto et al., 2001; Lau and Rosenthal, 2011; Metcalfe and Shimamura, 1994). On the other hand, many theoretical studies of neural circuits have come to the conclusion that probabilistic inference is the core computational process nervous systems perform (Knill and Pouget, 2004; Ma et al., 2006; Ma and Jazayeri, 2014; Rao, 2004; Rao et al., 2002; Zemel et al., 1998). Bayesian theories have been used to account for how brains can perform sensory processing (Beck et al., 2008; Ernst and Bülthoff, 2004; Graf et al., 2005) and motor control (Körding and Wolpert, 2004, 2006). From this perspective assigning a confidence level to a belief is an elementary computational task that lies at the core of statistical inference. Therefore understanding confidence in beliefs requires solid mathematical foundations, whether studied from the perspective of psychological, neural or machine learning research.

Previous theoretical studies have offered a number of different approaches to understand the statistical and algorithmic issues involved in computing and deploying decision confidence. For instance, a signal detection theory framework is often employed for probing decisions under uncertainty, and can provide a strong basis for understanding decision confidence as well (Fleming and Dolan, 2010; Kepecs et al., 2008; Ma, 2010; Maniscalco and Lau, 2012; Ratcliff and Starns, 2009). Sequential sampling models have been used to understand how decisions are reached based on noisy evidence across time (Bogacz et al., 2006). These can be readily extended with a computation of confidence (Drugowitsch et al., 2014; Pleskac and Busemeyer, 2010; Schustek and Moreno-Bote, 2014; Vickers, 1979). Perhaps the most intuitive extension is within the race model framework, where the difference between decision variables for the winning

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and losing races provides an estimate of confidence (Kepecs et al., 2008; Merkle and Van Zandt, 2006; Moreno-Bote, 2010; Vickers, 1979; Zylberberg et al., 2012). Mechanistically, neural network models based on attractor dynamics have also been used to study how confidence can be computed by neural circuits (Insabato et al., 2010; Rolls et al., 2010).

Such computational models have also helped to interpret experimental studies on the neural basis of decision confidence. However, it remains unclear how one could identify a confidence computation among mixed signals acquired from the brain, or how confidence in non-human animals can be quantified without verbal reports of their subjective feelings. What would a neural or behavioral implementation of confidence look like with respect to other observable measures of a decision? To resolve this quandary, previous studies employed quantitative models that could provide a formal prediction for what a representation of the internal variable of "confidence" would look like in terms of observable and quantifiable parameters. For instance, the orbitofrontal cortex of rats, a region implicated in the prediction of outcomes, was found to carry neural signals related to confidence (Kepecs et al., 2008). This was established by identifying unique signatures of confidence common to signal detection theory and the race model of decision-making. Similarly, signal detection theory predictions have been used to understand correlates of decision confidence in the dorsal pulvinar (Komura et al., 2013) and sequential sampling models in the parietal cortex (Kiani and Shadlen, 2009). Without such computational foundations, it would not be possible to identify and rigorously study representations of confidence in neurons.

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Beyond a description of how confidence could be computed, signal detection theory has also been used as a starting point for evaluating the metacognitive sensitivity of human confidence reports (Ferrell and McGoey, 1980; Higham and Arnold, 2007; Higham et al., 2009; Kunimoto et al., 2001; Lachman et al., 1979; Nelson, 1984). Indeed, even direct self-reports of the subjective sense of confidence require a computational framework to understand whether confidence reports are veridical, and the degree to which the use of evidence is optimal. For instance, in analogy with the *d'* measure (a classic measure describing how well an observer can distinguish two states of the world), "type-2" *d'* has been introduced to measure how well an observers' responses correspond to the objective state of the world (Fleming and Lau, 2014). Although this measure can be used to derive metacognitive sensitivity, it is based on the general signal detection theory assumption of Gaussian noise with equal variances. When these assumptions are violated, type-2 *d'* can be confounded by biases (Barrett et al., 2013; Evans and Azzopardi, 2007; Galvin et al., 2003; Kunimoto et al., 2001; Maniscalco and Lau, 2012).

Here we approached the well-studied topic of decision confidence from a mathematical statistics perspective. We had two main goals. First, compared to prior studies, we attempted to make as few assumptions as possible about the structure of noise and decision rule or the algorithm used for estimating confidence. Second, we approached the question of confidence from a psychophysical perspective so it may be useful for psychological and neural studies that often use perceptual uncertainty. The premise of our framework is a normative model of confidence which relates confidence to evidence through conditional probability (Kahneman and Tversky, 1972). While this premise is

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

widely accepted, we show that beyond calibration to outcome probabilities, it makes strong predictions about how a measure of confidence should relate to the discriminability of experimentally presented decision evidence.

We began from first principles in statistics by positing that confidence is a probability estimate describing a belief (Cox, 2006). Thus confidence can be related to the available evidence supporting the same belief through a conditional probability. As such, Bayes' rule provides a way to understand confidence in terms of quantifiable evidence (Ferrell and McGoey, 1980; Griffin and Tversky, 1992). Formally, decision confidence can be defined as a probability estimate that the chosen hypothesis is correct, given the available perceptual evidence – referred to as the percept. The difficulty with this definition of decision confidence is that it uses the percept, a variable internal to the decision maker. Therefore, it is unclear whether predictions are feasible without explicit assumptions about perception, how the internal percept is generated from the external stimulus. Such assumptions are generally used in the signal detection framework to keep the percept variable mathematically tractable. Here we show, however, that it is possible to analytically derive several novel predictions interrelating confidence with choice correctness and evidence discriminability with few or no assumptions about the percept distribution or about the transfer functions between stimulus, percept and choice.

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2 Results

From a statistical perspective, a decision process can be viewed as a hypothesis test that evaluates the outcome of a choice against a null hypothesis representing its collective alternatives. Statistical decision confidence can then be defined as a Bayesian posterior probability, which quantifies the degree of belief in the correctness of the chosen hypothesis. In this view, both choice and confidence depend on the quality and amount of evidence informing the particular choice. Therefore we mathematically formalized evidence discriminability based on ideas from psychophysics, as a way to measure the quality of evidence presented to a decision maker.

Based on these definitions, we derived four general properties of statistical decision confidence. First, confidence predicts accuracy: the level of confidence predicts the expected fraction of correct choices – as often intuitively posited. Second, confidence increases with the discriminability of presented evidence for correct choices, but counterintuitively, for incorrect choices, confidence decreases with increasing evidence-discriminability. Third, when presented with a zero-discriminability choice (i.e. an equal amount of evidence supporting each hypothesis, implying chance decision accuracy), the mean decision confidence is precisely 0.75. Fourth, while evidence discriminability itself predicts accuracy (a property referred to as the psychometric function), confidence provides further information improving the prediction of accuracy for any given level of discriminability.

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.1 Defining statistical decision confidence

To provide the most general statistical model of a decision process, we define all relevant components (stimulus, percept, choice, confidence) as random variables and the functions that link them (perception, decision) as probabilistic functions. This way the theory presented below applies to both stochastic and deterministic decision rules potentially involving multidimensional stimuli and multiple choices. Decisions are based on an internal variable (decision variable or percept, $\hat{D}$), which is the decision makers estimate of a corresponding external variable (stimulus or evidence, D).

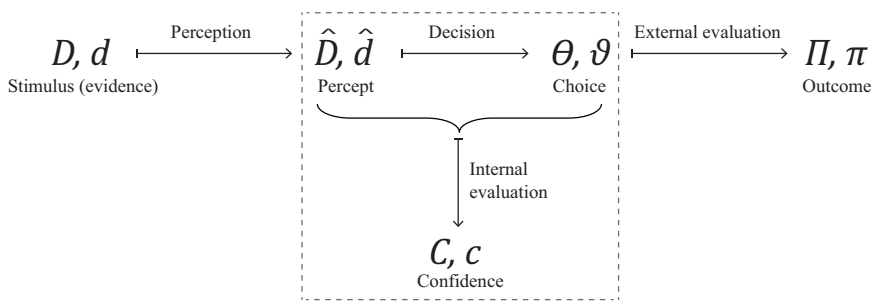

Figure 1: **A framework for statistical decision confidence.** A stochastic framework of perceptual decision making can be formalized by introducing a small set of random variables. Random variables are denoted by capital letters, and their realizations in lower case.

Definition 1. Let us denote the external variable D and realizations of this random variable d (referred to as *evidence*). Let us denote the corresponding internal variable $\hat{D}$ and realizations of this random variable $\hat{d}$ (referred to as *percept*; often referred to elsewhere as the *decision variable*). We define another random variable called the *choice*, denoted by θ (realizations denoted by ϑ). The choice is a probabilistic function of the percept:

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$\vartheta = \theta(\hat{d})$.

The choice can be evaluated in terms of a hypothesis testing problem:

null-hypothesis (H_0): the choice $\vartheta = \theta(\hat{d})$ is incorrect;

alternative hypothesis (H_1) : the choice $\vartheta = \theta(\hat{d})$ is correct.

Thus, the choice is designated *correct* if the alternative hypothesis is true and *incorrect* otherwise. The evaluation can equivalently be defined as a binary random variable (*outcome*, Π) that is a probabilistic function of choice. Next, confidence (c) can be defined as the probability of the alternative hypothesis being true (i.e. $\Pi(\theta) = 1$) provided the percept and the choice.

Definition 2. Define *confidence* as

$$c = P(H_1|\hat{d}, \vartheta). \tag{1}$$

Equivalently,

$$c = P(\Pi(\theta) = 1|\hat{d}, \vartheta). \tag{2}$$

As previously, the random variable will be denoted by C and its realizations by c. Note that for deterministic choice models, $\hat{d}$ determins ϑ, so $c = P(H_1|\hat{d})$. We can now define a function that determines confidence from percept and choice.

Definition 3. Define the function $\xi : \mathcal{R}(\hat{D}) \times \mathcal{R}(\theta) \to [0, 1]$ as

$$\xi(\hat{d}, \vartheta) = P(H_1|\hat{d}, \vartheta) = P(\Pi(\theta) = 1|\hat{d}, \vartheta). \tag{3}$$

where $\mathcal{R}(\hat{D})$ denotes *percept space* and $\mathcal{R}(\theta)$ denotes the range of all possible choices (i.e., the *choice space*).

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.2 Choice accuracy equals statistical decision confidence

Intuitively, confidence, being defined as an estimate of choice correctness, should predict the expected outcome. We provide a formal treatment of the relationship between confidence and accuracy below.

Definition 4. *Accuracy* is the expected proportion of correct choices:

$$A = E[\Pi(\theta)]. \tag{4}$$

We seek to determine the following function: $f : [0, 1] \to [0, 1]$, $f : c \mapsto A_c$, where A_c is the accuracy for choices with a given confidence. Our claim is that this function is the identity.

Theorem 1. Accuracy equals confidence:

$$A_c = c. \tag{5}$$

Proof. For every given value of confidence, there is a set of percept-choice pairs leading to the same confidence value: let us denote the image of c by ξ^{-1} as $\{(\hat{d}_i, \vartheta_i)\}_{i \in I}$, the set of percept-choice pairs mapping onto c. Let us first assume that I is a countable set. Accuracy for confidence c is determined by the probability of a correct choice if $C = c$ over the probability of encountering the confidence level of c (that is, $P(C = c)$):

$$A_c = \frac{\sum_{i \in I} P(H_1, \hat{d}_i, \vartheta_i)}{\sum_{i \in I} P(\hat{d}_i, \vartheta_i)}.$$

From the definition of joint probability,

$$A_c = \frac{\sum_{i \in I} P(H_1, \hat{d}_i, \vartheta_i)}{\sum_{i \in I} P(\hat{d}_i, \vartheta_i)} = \frac{\sum_{i \in I} P(H_1 | \hat{d}_i, \vartheta_i) \cdot P(\hat{d}_i, \vartheta_i)}{\sum_{i \in I} P(\hat{d}_i, \vartheta_i)}.$$

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

As we know that $\forall i \in I : P(H_1|\hat{d}_i, \vartheta_i) = c$,

$$A_c = \frac{\sum_{i\in I} P(H_1|\hat{d}_i, \vartheta_i) \cdot P(\hat{d}_i, \vartheta_i)}{\sum_{i\in I} P(\hat{d}_i, \vartheta_i)} = \frac{\sum_{i\in I} c \cdot P(\hat{d}_i, \vartheta_i)}{\sum_{i\in I} P(\hat{d}_i, \vartheta_i)} = \frac{c \cdot \sum_{i\in I} P(\hat{d}_i, \vartheta_i)}{\sum_{i\in I} P(\hat{d}_i, \vartheta_i)} = c.$$

However, I is not necessarily a countable set. We can re-write the equations in continuous form to apply to any set as follows.

$$A_c = \frac{\int_{\Pi=1} \iint_{(\hat{D},\theta)\in\xi^{-1}(c)} f_{\Pi,\hat{D},\theta}(\pi, \hat{d}, \vartheta)\mathrm{d}\vartheta\mathrm{d}\hat{d}\mathrm{d}\pi}{\iint_{(\hat{D},\theta)\in\xi^{-1}(c)} f_{\hat{D},\theta}(\hat{d}, \vartheta)\mathrm{d}\vartheta\mathrm{d}\hat{d}}$$

Here Π is a random variable that is 1 if the choice is correct and 0 otherwise (*outcome*, see above).

$$\begin{aligned}
A_c &= \frac{\int_{\Pi=1} \iint_{(\hat{D},\theta)\in\xi^{-1}(c)} f_{\Pi,\hat{D},\theta}(\pi, \hat{d}, \vartheta)\mathrm{d}\vartheta\mathrm{d}\hat{d}\mathrm{d}\pi}{\iint_{(\hat{D},\theta)\in\xi^{-1}(c)} f_{\hat{D},\theta}(\hat{d}, \vartheta)\mathrm{d}\vartheta\mathrm{d}\hat{d}} \\
&= \frac{\int_{\Pi=1} \iint_{(\hat{D},\theta)\in\xi^{-1}(c)} f_{\Pi}(\pi|\hat{D} = \hat{d}, \theta = \vartheta) \cdot f_{\hat{D},\theta}(\hat{d}, \vartheta)\mathrm{d}\pi\mathrm{d}\vartheta\mathrm{d}\hat{d}}{\iint_{(\hat{D},\theta)\in\xi^{-1}(c)} f_{\hat{D},\theta}(\hat{d}, \vartheta)\mathrm{d}\vartheta\mathrm{d}\hat{d}} \\
&= \frac{\int_{\Pi=1} \iint_{(\hat{D},\theta)\in\xi^{-1}(c)} c \cdot f_{\hat{D},\theta}(\hat{d}, \vartheta)\mathrm{d}\pi\mathrm{d}\vartheta\mathrm{d}\hat{d}}{\iint_{(\hat{D},\theta)\in\xi^{-1}(c)} f_{\hat{D},\theta}(\hat{d}, \vartheta)\mathrm{d}\vartheta\mathrm{d}\hat{d}} \\
&= \frac{c \cdot \iint_{(\hat{D},\theta)\in\xi^{-1}(c)} f_{\hat{D},\theta}(\hat{d}, \vartheta)\mathrm{d}\vartheta\mathrm{d}\hat{d}}{\iint_{(\hat{D},\theta)\in\xi^{-1}(c)} f_{\hat{D},\theta}(\hat{d}, \vartheta)\mathrm{d}\vartheta\mathrm{d}\hat{d}} \\
&= c.
\end{aligned}$$

□

Note that these considerations about confidence do not depend on a particular theory of perception, that is, the function mapping from the external variable (stimulus) onto the internal percept: $D \mapsto \hat{D}$. Furthermore, the derivation also does not depend on a particular theory of decision, that is, the function between the percept and the choice: $\hat{D} \mapsto \theta$. This includes both deterministic and stochastic decision models, the latter referring to models where a percept does not uniquely determine a choice. In

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

case of deterministic decision models, the percept unequivocally determines the choice, thus in the equations we could drop the choice from the inverse picture of confidence, taking only the percept into account: $\{\hat{d}_i\}_{i\in I}$ instead of $\{(\hat{d}_i, \vartheta_i)\}_{i\in I}$. However, as this simplified version would not include stochastic decision models, we chose to adhere to the general formalization.

Another notable aspect of this derivation is that there is no need for a relation to be defined on the percept space. However, if the choice is fixed (or determined by the percept, as in deterministic decision models), confidence defines a natural relation on percepts by ξ. More precisely, the order relation on confidence values can be pulled back to the percept space by taking $\xi^{-1}(c)$ and restricting it to a particular choice.

2.3 Confidence increases with increasing evidence discriminability for correct choices and decreases for incorrect choices

Psychophysical studies requires to measure decision performance at varying levels of decision difficulty. This necessitates the quantification of the decision difficulty axis, along which the proportion of correct choices can then be measured. Such interrelations, termed psychometric functions, provide a good handle on behavioral performance allowing the detection of subtle changes in behavior. However, there is no single way of grading choice difficulty, resulting in a broad variety of such measures, which complicates the theoretical treatment of psychometric functions. Therefore we define evidence discriminability by its property of measuring difficulty as a class of functions in order to provide a general treatment of the interrelations of choice difficulty and confidence.

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Definition 5. Define *evidence discriminability* as a (deterministic) function of the evidence distribution:

$$\Delta = \Delta(P(D)) \tag{6}$$

The evidence discriminability function has to fulfill the following property:

$$\begin{aligned} \Delta(P_1(D)) > \Delta(P_2(D)) \quad &\Longleftrightarrow \quad P(H_1|P_1(D)) > P(H_1|P_2(D)) \\ &\Longleftrightarrow \quad P(\Pi(\theta) = 1|P_1(D)) > P(\Pi(\theta) = 1|P_2(D)) \\ &\Longleftrightarrow \quad E(\Pi(\theta)|P_1(D)) > E(\Pi(\theta)|P_2(D)), \end{aligned} \tag{7}$$

that is, higher discriminability should be equivalent to greater expected outcome (higher probability of correct choices). Any monotonically increasing function of expected outcome satisfies this criterion and can serve as evidence discriminability.

Having defined evidence discriminability, we can now examine how confidence changes with evidence discriminability separately for correct and incorrect choices. We show below that while confidence increases with increasing evidence discriminability for correct choices, it counterintuitively decreases for incorrect choices.

Theorem 2. Let us assume that

- ξ is independent of evidence discriminability;
- the distribution of percepts changes monotonically with evidence discriminability.

Under these assumptions, confidence increases for correct choices and decreases for incorrect choices with increasing evidence discriminability.

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Proof. We begin with the somewhat counterintuitive claim regarding the incorrect choices. Let us first examine the two assumptions in more detail.

The first assumption postulates that the function from percept-choice pairs to confidence does not change with evidence discriminability. Thus, whenever we calculate expected value of confidence over a percept distribution, only the percept distributions will depend on evidence discriminability.

For incorrect choices, the second assumption means that with increasing evidence discriminability, the relative weight of low-confidence percepts increases while the relative weight of high-confidence percepts decreases in the percept distribution. Note that low-confidence and high-confidence percepts are defined here through the relation imposed by ξ on the percepts (see our remark at the end of the previous section). As a trivial consequence of this definition, confidence changes monotonically along low- and high-confidence percepts.

Let us consider two different levels of evidence discriminability ($\Delta_1 < \Delta_2$), with corresponding distributions of percept restricted to incorrect choices P (Δ_1, low evidence discriminability, i.e. 'difficult choice') and Q (Δ_2, high evidence discriminability, i.e. 'easy choice'). It is sufficient to show that the expected value of confidence is larger for Δ_1 than for Δ_2:

$$\int_0^1 c \cdot p(c)\mathrm{d}c > \int_0^1 c \cdot q(c)\mathrm{d}c$$

where p and q denotes the probability density functions corresponding to P and Q, respectively. Note that $p(c)$ can be thought of as the probability of the picture of c by ξ^{-1} restricted to incorrect choices in the percept space.

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Equivalently,

$$\int_0^1 c \cdot [p(c) - q(c)]\mathrm{d}c > 0$$

Let $I_0 \subset [0,1]$ denote the interval where $p < q$, and $I_1 \subset [0,1]$ the complementary interval where $p \geq q$. The existence of these intervals is the consequence of the monotonicity assumption. Thus, there is a critical confidence value (denoted here by c_{crit}) for which $I_0 = [0, c_{crit}]$ and $I_1 = [c_{crit}, 1]$. We then re-write confidence as $c = c_{crit} - c'$ if $c < c_{crit}$ and $c = c_{crit} + c'$ if $c > c_{crit}$; thus, $c' > 0$ for both cases. Applying these notations,

$$\begin{aligned}
\int_0^1 c \cdot [p(c) - q(c)]\mathrm{d}c &= \int_0^{c_{crit}} c \cdot [p(c) - q(c)]\mathrm{d}c + \int_{c_{crit}}^1 c \cdot [p(c) - q(c)]\mathrm{d}c \\
&= \int_{c_{crit}}^0 (c_{cirt} - c') \cdot [p(c_{crit} - c') - q(c_{crit} - c')]\mathrm{d}c' + \\
&\quad \int_0^{1-c_{crit}} (c_{crit} + c') \cdot [p(c_{crit} + c') - q(c_{crit} + c')]\mathrm{d}c' \\
&= c_{crit} \cdot \Bigg(\int_{c_{crit}}^0 [p(c_{crit} - c') - q(c_{crit} - c')]\mathrm{d}c' + \\
&\quad \int_0^{1-c_{crit}} [p(c_{crit} + c') - q(c_{crit} + c')]\mathrm{d}c' \Bigg) + \\
&\quad \int_0^{1-c_{crit}} c' \cdot [p(c_{crit} + c') - q(c_{crit} + c')]\mathrm{d}c' - \\
&\quad \int_{c_{crit}}^0 c' \cdot [p(c_{crit} - c') - q(c_{crit} - c')]\mathrm{d}c' \\
&= c_{crit} \cdot \Bigg(\int_0^{c_{crit}} [p(c) - q(c)]\mathrm{d}c + \int_{c_{crit}}^1 [p(c) - q(c)]\mathrm{d}c \Bigg) + \\
&\quad \int_{c_{crit}}^1 (c - c_{crit}) \cdot [p(c) - q(c)]\mathrm{d}c - \int_0^{c_{crit}} (c_{crit} - c) \cdot [p(c) - q(c)]\mathrm{d}c
\end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$
\begin{aligned}
&= c_{crit} \cdot \left(\int_0^1 [p(c) - q(c)] \mathrm{d}c \right) + \\
&\int_{c_{crit}}^1 (c - c_{crit}) \cdot [p(c) - q(c)] \mathrm{d}c + \int_0^{c_{crit}} (c_{crit} - c) \cdot [q(c) - p(c)] \mathrm{d}c \\
&> 0
\end{aligned}
$$

In the last step, the first term is 0, since

$$\int_0^1 [p(c) - q(c)] \mathrm{d}c = \int_0^1 p(c) \mathrm{d}c - \int_0^1 q(c) \mathrm{d}c = 1 - 1 = 0$$

The second term is positive, since $c - c_{crit}$ is positive on $c \in [c_{crit}, 1]$ and the probability density functions are evaluated on I_1, where $p \geq q$. Finally, the third term is also positive, because $c_{crit} - c$ is positive on $c \in [0, c_{crit}]$ and the probability density functions are evaluated on I_0, where $q > p$. As a consequence, the sum is positive, which completes the proof for incorrect choices.

For correct choices, high-confidence percepts are increasingly more likely with increasing evidence discriminability, thus present an opposite pattern compared to incorrect choices. Therefore, a symmetric derivation proves the increase of confidence with increasing evidence discriminability for correct choices. □

The assumption that ξ is independent of evidence discriminability is necessary for this derivation. If confidence values are defined based on distributions reflecting varying values of evidence discriminability, then the function between the percept and confidence (ξ) will differ according to evidence discriminability, thus the above proof does not apply. Furthermore, the expected value of confidence cannot decrease with increasing evidence discriminability for incorrect choices: for the lowest levels of dis-

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

criminability, when the outcome is at chance level, confidence will fall to its lowest possible value, reflecting equal probabilities of the null and alternative hypothesis regardless of the percept. This represents situations in which the decision-maker is provided with information about evidence discriminability, e.g. by grouping decisions of similar discriminabilities, providing an opportunity to learn about evidence discriminability and defining confidence accordingly. Thus, the above theorem only applies when changing confidence levels based on knowledge of evidence discriminability is prevented, e.g. by randomizing the order of choices with different discriminability levels.

2.4 Confidence predicts outcome beyond evidence discriminability

Psychometric functions reveal accuracy for any given level of evidence discriminability. While confidence also changes with evidence discriminability, it is not obvious whether it carries additional information allowing better prediction of outcome for a given level of evidence discriminability. Below we show that it does.

Theorem 3. For any given evidence discriminability, accuracy for low confidence choices is not larger than that of high confidence choices (splitting the confidence distribution at any particular value). A strict inequality holds in all cases when accuracy is dependent on the percept.

Proof. Let us take the set of low-confidence percept-choice pairs corresponding to the low confidence choices by ξ^{-1}, and similarly, the set of high-confidence percept-choice

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

pairs corresponding to the high confidence choices. By the definition of confidence (*Definition 2* in *Section 2.1*), low-confidence percept-choice pairs cannot have higher accuracy than the high-confidence percept-choice pairs. If all percepts are associated with the same accuracy (either when the percept does not carry information about the hypotheses of choice, or when the percept determines the correct choice with a probability of one), the two accuracies are equal. Otherwise, the two accuracies should necessarily differ, in which case the strict inequality holds. □

2.5 The average confidence in neutral evidence

Next, we examine the average confidence at neutral evidence, i.e. evidence carrying no information about the correct choice, for one-dimensional variables.

Theorem 4. Assuming

- the percept is determined by a symmetric distribution centered on the evidence ('symmetric noise model'),
- the evidence is distributed uniformly over the evidence space, and
- the choice is deterministic,

the average confidence for neutral evidence is precisely 0.75.

Proof. We first prove the following lemma.

Lemma 5. Integrating the product of the probability density function and the distribution function of any probability distribution symmetric to zero over the positive half-line

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

results in 3/8:

$$\int_{0}^{\infty} f(t)F(t)\mathrm{d}t = \frac{3}{8}. \tag{8}$$

Proof. $\forall K\ -\infty < K < \infty$,

$$\begin{aligned}
\int_{t=-\infty}^{K} f(t)F(t)\mathrm{d}t &= \int_{t=-\infty}^{K} f(t) \int_{x=-\infty}^{t} f(x)\mathrm{d}x\mathrm{d}t \\
&= \int_{t=-\infty}^{K} \int_{x=-\infty}^{t} f(t)f(x)\mathrm{d}x\mathrm{d}t \\
&= \frac{1}{2}2 \int_{t=-\infty}^{K} \int_{x=-\infty}^{t} f(t)f(x)\mathrm{d}x\mathrm{d}t \\
&= \frac{1}{2}\Big[\int_{t=-\infty}^{K} \int_{x=-\infty}^{t} f(t)f(x)\mathrm{d}x\mathrm{d}t + \int_{t=-\infty}^{K} \int_{x=-\infty}^{t} f(t)f(x)\mathrm{d}x\mathrm{d}t \Big] \\
&= \frac{1}{2}\Big[\int_{t=-\infty}^{K} \int_{x=-\infty}^{t} f(t)f(x)\mathrm{d}x\mathrm{d}t + \int_{x=-\infty}^{K} \int_{t=x}^{K} f(x)f(t)\mathrm{d}t\mathrm{d}x \Big] \\
&= \frac{1}{2}\Big[\int_{t=-\infty}^{K} \int_{x=-\infty}^{t} f(t)f(x)\mathrm{d}x\mathrm{d}t + \int_{t=-\infty}^{K} \int_{x=t}^{K} f(t)f(x)\mathrm{d}x\mathrm{d}t \Big] \\
&= \frac{1}{2}\Big[\int_{t=-\infty}^{K} f(t) \int_{x=-\infty}^{t} f(x)\mathrm{d}x\mathrm{d}t + \int_{t=-\infty}^{K} f(t) \int_{x=t}^{K} f(x)\mathrm{d}x\mathrm{d}t \Big] \\
&= \frac{1}{2}\Big[\int_{t=-\infty}^{K} f(t)\Big[\int_{x=-\infty}^{t} f(x)\mathrm{d}x + \int_{x=t}^{K} f(x)\mathrm{d}x \Big]\mathrm{d}t \Big] \\
&= \frac{1}{2}\Big[\int_{t=-\infty}^{K} f(t) \int_{x=-\infty}^{K} f(x)\mathrm{d}x\mathrm{d}t \Big] \\
&= \frac{1}{2}\Big[\int_{t=-\infty}^{K} f(t)\mathrm{d}t \int_{x=-\infty}^{K} f(x)\mathrm{d}x \Big] \\
&= \frac{1}{2}\Big[\int_{t=-\infty}^{K} f(t)\mathrm{d}t \Big]^2
\end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

using that

$$\{-\infty < t < K, -\infty < x < t\} \iff \{-\infty < x < K, x < t < K\}$$

for changing the integral boundaries and then swapping x and t in the second integral term. Applying the above equation, we can write

$$\begin{aligned}\int_0^\infty f(t)F(t)\mathrm{d}t &= \int_{-\infty}^\infty f(t)F(t)\mathrm{d}t - \int_{-\infty}^0 f(t)F(t)\mathrm{d}t \\ &= \frac{1}{2}\left[\int_{-\infty}^\infty f(t)\mathrm{d}t\right]^2 - \frac{1}{2}\left[\int_{-\infty}^0 f(t)\mathrm{d}t\right]^2 \\ &= \frac{1}{2}1^2 - \frac{1}{2}\left[\frac{1}{2}\right]^2 = \frac{3}{8}\end{aligned}$$

□

Proof of the theorem: Confidence for neutral evidence is determined by the percept corresponding to neutral evidence and the probability of being correct provided the percept and the choice. Thus, the average confidence for neutral evidence can be calculated by integrating over the distribution of percepts provided neutral evidence (indicated here by $d = 0$):

$$c_0 = \int_{-\infty}^\infty P(\hat{d}|d = 0)\xi(\hat{d}, \vartheta)\mathrm{d}\hat{d}$$

Since we assumed deterministic choice (third assumption), confidence is determined by the percept; therefore we can drop ϑ from the equation:

$$c_0 = \int_{-\infty}^\infty P(\hat{d}|d = 0)\xi(\hat{d})\mathrm{d}\hat{d}$$

Based on our first assumption, the percept is determined by a symmetric distribution around the evidence. Denote the density function of this symmetric ('noise') distribu-

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

tion f and its distribution function F. Since the percept distribution is symmetric,

$$c_0 = \int_{-\infty}^{\infty} P(\hat{d}|d=0)\xi(\hat{d})\mathrm{d}\hat{d} = 2\int_{0}^{\infty} P(\hat{d}|d=0)\xi(\hat{d})\mathrm{d}\hat{d}$$

As a consequence of the second assumption of uniform evidence distribution, $\xi(\hat{d}) = F(\hat{d})$ for $\hat{d} > 0$. Thus,

$$c_0 = 2\int_{0}^{\infty} P(\hat{d}|d=0)\xi(\hat{d})\mathrm{d}\hat{d} = 2\int_{0}^{\infty} f(\hat{d})F(\hat{d})\mathrm{d}\hat{d}$$

By applying the lemma,

$$c_0 = 2\int_{0}^{\infty} f(\hat{d})F(\hat{d})\mathrm{d}\hat{d} = 2\frac{3}{8} = \frac{3}{4}.$$

□

Note that one of the critical assumptions we made is that the evidence is distributed uniformly over the evidence space. While real-life scenarios may often represent non-uniform evidence distributions, this uniformity property holds approximately true for many psychophysics experiments using interleaved evidence strength. Therefore *Theorem 4* provides a quantitatively testable prediction about confidence reports in psychophysics experiments.

2.6 Monte Carlo simulations illustrating the signatures of decision confidence

To illustrate our theory we created a Monte Carlo simulation of the normative definition of confidence. For the simulation, we assumed that Gaussian noise ($\mu = 0, \sigma = 0.18$)

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

corrupts the external evidence, to generate an internal percept. We used a deterministic decision rule based on the sign of the percept. Thus, outcomes were correct if the sign of the evidence and percept matched. Confidence was calculated as fraction of correct trials for each percept based on *Definition 2*. This enabled us to explore predicted interrelationships between confidence, evidence discriminability and choice. *Figure 2A* shows that confidence predicts the mean choice accuracy (*Theorem 1*). *Figure 2B* demonstrates that mean confidence for a given level of evidence discriminability increases for correct and decreases for incorrect choices (*Theorem 2*), and that the mean confidence for neutral evidence is 0.75 (*Theorem 4*). *Figure 2C* illustrates that for each given level of evidence discriminability, accuracy for high confidence choices is greater than for low confidence choices (*Theorem 3*). Taken together these plots illustrate four signatures of decision confidence in terms of externally quantifiable variables that can be experimentally examined.

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

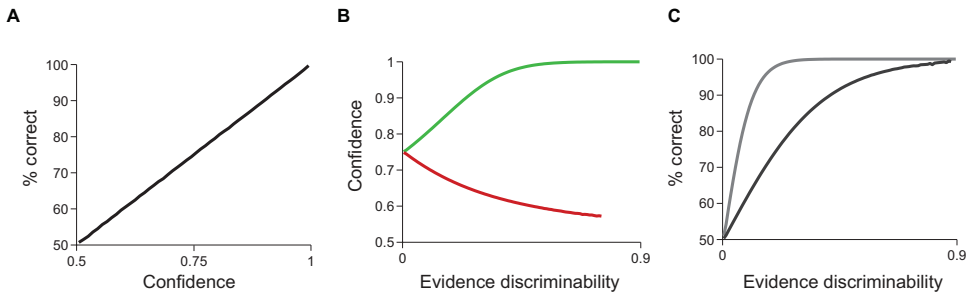

Figure 2: **The normative model of confidence predicts specific interrelationships between evidence, outcome and confidence.** **(A-C)** Monte Carlo simulations of the normative model (10 billion trials). Bins with fewer than 100 simulation data points were omitted. **(A)** Confidence equals accuracy. **(B)** Average confidence increases with evidence discriminability from 0.75 for correct choices and decreases for errors. **(C)** Conditioning on high or low confidence (split at $c = 0.8$) segregates psychometric performance.

2.7 Relating P-values to confidence

Above we derived properties of statistical confidence based on the definition of decision confidence as a Bayesian posterior probability. We next sought to demonstrate their generality by testing their validity on confidence values produced by other statistical approaches. First, we constructed a simulation to test the properties of p-values produced by a common statistical test for evaluating a choice between two hypotheses. First, we examined the one-sided, two-sample Students t-test (*Figure 3A-C*). Samples of 20 measurements were drawn from two Gaussian distributions on each simulation trial, where the simulated task was to identify which underlying distribution had a larger mean. To create graded discriminability, we varied the distance between the means from -0.5 to

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

0.5 with uniform probability. A simulation trial was designated as "correct" if the mean of the 20 samples drawn from the distribution with the higher mean was higher than the mean of the 20 samples drawn from the distribution with the lower mean. We computed the p-value for each trial using a one sided two-sample t-test to provide a measure of statistical confidence $(1 - p)$ in the chosen response. Thus each simulation trial yielded an outcome (correct or error) and a measure of statistical confidence. Second, we also performed simulations of a bootstrap test (Efron and Tibshirani, 1993), which does not depend on a Gaussian assumption about the underlying distributions (*Figure 3D-F*). Exponential sample distributions were used. Offsets for the population means were uniform, ranging between 0 and 1, and the bootstrap sample size was 1000. As shown in *Figure 3*, the p-values derived from a t-test and a two-sample bootstrap test for difference between means reveal the same pattern of interrelationships we derived from the Bayesian confidence definition. Thus, the predictions we derived for statistical decision confidence are valid across different statistical approaches: Bayesian, frequentist and bootstrap statistics.

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

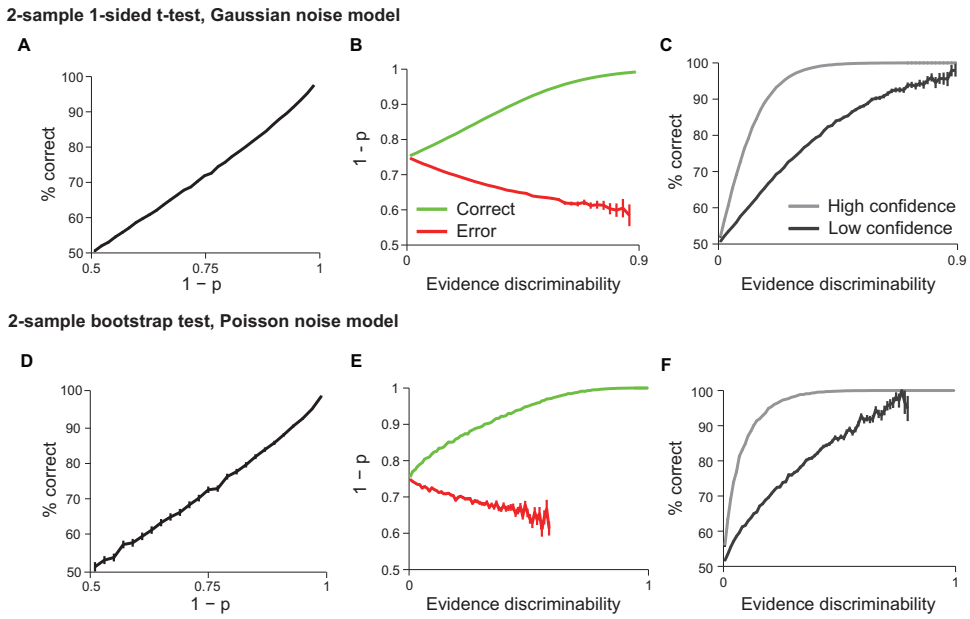

Figure 3: **Two statistical tests reproduce patterns predicted by the normative model.** **(A-C)** A simulation of 10 million trials evaluated the one-sided, two sample Students t-test p-value with respect to accuracy and evidence discriminability. Since p indicates uncertainty, axes show $1 - p$ to indicate confidence. **(A)** $1 - p$ is positively correlated with accuracy. **(B)** $1 - p$ is monotonically increasing with evidence discriminability for correct trials and decreasing for error trials. **(C)** P-values contain information about outcome even at fixed evidence discriminability. **(D-F)** A simulation of the p-value in a one-sided bootstrap test for an ordinal relationship between two means, using exponential distributions.

3 Discussion

We presented a normative statistical framework that enables comparisons of statistical decision confidence with confidence measures in other domains. Unlike signal detection

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

theory and other algorithmic frameworks that simulate confidence judgments based on assumptions about the underlying evidence distributions, we show that a strict analytical treatment is possible in a distribution-free manner.

We analytically derived a set of properties of confidence defined as the Bayesian posterior probability of a chosen hypothesis being correct. First, confidence predicts accuracy: assuming that the distribution of possible evidence quality is known (i.e. based on experience), then the level of confidence predicts the expected fraction of correct choices. This property corresponds most directly to the intuitive notion of confidence as a graded forecast about accuracy. Second, mean confidence for a given level of external evidence is larger for correct than incorrect choices and in fact varies with an opposite sign with evidence discriminability for correct vs. incorrect choices. Specifically, mean confidence levels increase with the ease of discriminability for correct choices, but counterintuitively, confidence decreases with increasing evidence-discriminability for incorrect choices. This surprising dissociation is a consequence of the differences in the distributions of conditional percepts between correct and incorrect choices. Third, and perhaps most surprisingly, when presented with an equal amount of evidence supporting each hypothesis, in other words a non-discriminable choice that will lead to chance accuracy, the mean decision confidence is much greater than chance – precisely 0.75. Fourth, while the psychometric function defines the average choice accuracy for a given level of external evidence, knowledge of confidence provides further information improving the prediction of accuracy for any given level of discriminability.

These four properties are useful for interpreting both behavioral and physiolog-

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ical experiments on decision confidence. Behaviorally, our framework makes it clear that even statistically optimal confidence reports can appear to show systematic miscalibration. This mismatch between confidence reports and accuracy is most dramatically illustrated by the 0.75 mean confidence for neutral evidence that produces chance accuracy behaviorally (0.5). As our framework makes it clear this apparent miscalibration does not imply imperfect prediction of accuracy – rather it is a straightforward consequence of conditioning confidence reports on external variables of the task (e.g,. stimulus difficulty) that are not available to the decision maker. This property of statistical confidence carries important implications for the interpretation of studies demonstrating overconfidence in low discriminability and under-confidence in high discriminability conditions – a controversial phenomenon termed the "hard-easy effect" (Drugowitsch et al., 2014; Ferrell, 1995; Harvey, 1997; Juslin et al., 2000; Merkle, 2009; Moore and Healy, 2008). More generally one has to be careful when analyzing behavior or neural activity by conditioning on external variables not available to the decision maker rather than the internal representations that they are based on. When internal representations are examined as a function of external variables a computational theory is needed to understand how observables conditioned on the external variables is linked to the internal representations. Therefore, rather than revealing miscalibration, conditioning on external variables can be used to test signatures of decision confidence we derived (*Figure 2*) and will be valuable in interpreting putative confidence-related neural activity as well (Kepecs et al., 2008; Komura et al., 2013; Kiani and Shadlen, 2009).

The framework we presented can be interpreted as a prescriptive model, describing how the computation of confidence ought to be done. In this sense it is useful for

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

describing what a neural representation of confidence or its behavioral report should look like. Beyond this, we expect that our mathematical framework will serve as a departure point for quantitatively studying the contribution of confidence to different behaviors, and identifying confidence variables in other domains.

References

Barrett, A. B., Dienes, Z., and Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological methods*, 18(4):535–52.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., and Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron*, 60(6):1142–1152.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4):700.

Charles, L., Van Opstal, F., Marti, S., and Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage*, 73(0):80–94.

Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press.

Drugowitsch, J., Moreno-Bote, R., and Pouget, A. (2014). Relation between Belief and Performance in Perceptual Decision Making. *PloS one*, 9(5):e96511.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*, volume 57. CRC press.

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ernst, M. O. and Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in cognitive sciences*, 8(4):162–9.

Evans, S. and Azzopardi, P. (2007). Evaluation of a'bias-free'measure of awareness. *Spatial vision*, 20(1):61–78.

Ferrell, W. R. (1995). A model for realism of confidence judgments: implications for underconfidence in sensory discrimination. *Perception & psychophysics*, 57(2):246–54; discussion 255–9.

Ferrell, W. R. and McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance*, 26(1):32–53.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitivedevelopmental inquiry. *American psychologist*, 34(10):906.

Fleming, S. M. and Dolan, R. J. (2010). Effects of loss aversion on post-decision wagering: Implications for measures of awareness. *Consciousness and cognition*, 19(1):352–363.

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

Galvin, S. J., Podd, J. V., Drga, V., and Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychon Bull Rev*, 10(4):843–876.

Graf, E. W., Warren, P. A., and Maloney, L. T. (2005). Explicit estimation of visual uncertainty in human motion processing. *Vision research*, 45(24):3050–9.

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Griffin, D. and Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive psychology*, 24(3):411–435.

Harvey, N. (1997). Confidence in judgment. *Trends in cognitive sciences*, 1(2):78–82.

Higham, P. A. and Arnold, M. M. (2007). Beyond reliability and validity: The role of metacognition in psychological testing. *New developments in psychological testing*, pages 139–162.

Higham, P. A., Perfect, T. J., and Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(1):57.

Insabato, A., Pannunzi, M., Rolls, E. T., and Deco, G. (2010). Confidence-related decision making. *Journal of neurophysiology*, 104(1):539–547.

Juslin, P., Winman, A., and Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hardeasy effect. *Psychological Review*, 107(2):384.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719.

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):749–55.

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

Körding, K. P. and Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in cognitive sciences*, 10(7):319–26.

Kunimoto, C., Miller, J., and Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and cognition*, 10(3):294–340.

Lachman, J. L., Lachman, R., and Thronesbery, C. (1979). Metamemory through the adult life span. *Developmental Psychology*, 15(5):543.

Lau, H. and Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in cognitive sciences*, 15(8):365–373.

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

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–8.

Ma, W. J. and Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual review of neuroscience*, 37:205–20.

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

Merkle, E. C. (2009). The disutility of the hard-easy effect in choice confidence. *Psychonomic bulletin & review*, 16(1):204–213.

Merkle, E. C. and Van Zandt, T. (2006). An application of the Poisson race model to confidence calibration. *Journal of Experimental Psychology: General*, 135(3):391.

Metcalfe, J. E. and Shimamura, A. P. (1994). *Metacognition: Knowing about knowing*. The MIT Press.

Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2):502.

Moreno-Bote, R. (2010). Decision confidence and uncertainty in diffusion models with partially correlated neuronal integrators. *Neural computation*, 22(7):1786–1811.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95(1):109.

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.

Rao, R. P. N. (2004). Bayesian computation in recurrent neural circuits. *Neural computation*, 16(1):1–38.

Rao, R. P. N., Olshausen, B. A., and Lewicki, M. S. (2002). *Probabilistic models of the brain: Perception and neural function*. Mit Press.

bioRxiv preprint doi: https://doi.org/10.1101/017400; this version posted April 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ratcliff, R. and Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological review*, 116(1):59–83.

Rolls, E. T., Grabenhorst, F., and Deco, G. (2010). Decision-making, errors, and confidence in the brain. *Journal of neurophysiology*, 104(5):2359–74.

Schustek, P. and Moreno-Bote, R. (2014). A theory of decision-making using diffusion-to-bound models: choice, reaction-time and confidence. *BMC Neuroscience*, 15(Suppl 1):P88.

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

Zemel, R. S., Dayan, P., and Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural computation*, 10(2):403–30.

Zylberberg, A., Barttfeld, P., and Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, 6.