Modeling perceptual confidence and the confidence forced-choice paradigm.

Pascal Mamassian, Vincent de Gardelle

▶ **To cite this version:**

Pascal Mamassian, Vincent de Gardelle. Modeling perceptual confidence and the confidence forced-choice paradigm.. 2021. hal-03452165

HAL Id: hal-03452165

https://cnrs.hal.science/hal-03452165v1

Preprint submitted on 26 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Modelling Perceptual Confidence and the Confidence Forced-Choice Paradigm

Pascal Mamassian[1] and Vincent de Gardelle[2]

(1) Laboratoire des systèmes perceptifs, Département d'études cognitives, École normale supérieure, PSL University, CNRS, Paris, France
(2) CNRS and Paris School of Economics, Paris, France

This is a preprint of the paper:

Mamassian, P., & de Gardelle, V. (2021). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review.* Advance online publication. https://doi.org/10.1037/rev0000312

Correspondence should be addressed to:
Pascal Mamassian, Laboratoire des Systèmes Perceptifs (CNRS UMR 8248)
Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France
Email: <pascal.mamassian@ens.fr>

Abstract

Perceptual confidence is an evaluation of the validity of our perceptual decisions. We present here a complete generative model that describes how confidence judgments result from some confidence evidence. The model that generates confidence evidence has two main parameters, confidence noise and confidence boost. Confidence noise reduces the sensitivity to the confidence evidence, and confidence boost accounts for information used for confidence judgment which was not used for the perceptual decision. The opposite effect of these two parameters creates a problem of confidence parameters indeterminacy, where the confidence in a perceptual decision is the same in spite of differences in confidence noise and confidence boost. When confidence is estimated for multiple stimulus strengths, both of these parameters can be recovered, thus allowing us to estimate whether confidence is generated using the same primary information that was used for the perceptual decision or some secondary information. We also describe a novel measure of confidence efficiency relative to the ideal confidence observer, as well as the estimate of one type of confidence bias. Finally, we apply the model to the confidence forced-choice paradigm, a paradigm that provides objective estimates of confidence, and we discuss how each parameter of the model can be recovered using this paradigm.

Keywords: meta-perception, visual confidence, modelling, efficiency, confidence forced-choice

1. Introduction

Metacognition is the ability of individuals to monitor and regulate their own cognitive processes (Nelson & Narens, 1990). This ability is referred to as meta-perception when considering the monitoring and regulation of perceptual processes and decisions (Mamassian, 2020). When making a choice, a key expression of metacognition is the confidence associated with the decision. Correctly inferring our own level of performance is clearly important for an individual, as confidence might be used to control the amount of sensory information necessary to commit to a perceptual decision (Balsdon, Wyart, & Mamassian, 2020), regulate learning (e.g. Hainguerlot, Vergnaud, & de Gardelle, 2018), allocate resources to a particular task (e.g. van den Berg et al., 2016), compare different tasks (de Gardelle & Mamassian, 2014) and prioritize them (Aguilar-Lleyda, Lemarchand, & de Gardelle, 2020). Perceptual confidence, and more broadly metacognition, has been extensively reviewed elsewhere (e.g. Fleming et al., 2012; Yeung & Summerfield, 2012; Meyniel, Sigman, & Mainen, 2015; Mamassian, 2016; Pouget et al., 2016). In this article, we develop a theoretical framework to characterize how individuals make confidence judgments about their perceptual decisions.

One issue of primary importance in meta-perception is whether confidence judgments are based on the same information as that used for the perceptual decisions. Here, to clearly separate perceptual decisions from confidence judgments, we formally distinguish the evidence underlying these two computations, calling *sensory evidence* the basis of our perceptual decisions and *confidence evidence* the basis of our confidence judgments. A similar distinction was made in meta-memory (Jang, Wallsten, & Huber, 2012) and meta-perception (Fleming & Daw, 2017). In addition, when considering confidence evidence, we further distinguish two components, calling *primary* confidence evidence the information that was used for the perceptual decisions, and *secondary* confidence evidence any other information (i.e. information contributing to confidence evidence but not to the perceptual decisions). Indeed, even though confidence is an evaluation of the validity of our perceptual decisions, it is plausible that the computation of confidence involves some information that is processed in parallel to (e.g. Fleming & Daw, 2017) or after (e.g. Pleskac & Busemeyer, 2010) the perceptual decision. The difficulty in establishing the extent to which confidence is relying on secondary evidence is that there are other factors that affect the quality of confidence judgments. In particular, the computation of confidence might rest on perceptual information that has been degraded by some form of confidence noise (e.g. Bang et al., 2019). Therefore, it is important to have a good theoretical framework within which the different factors that contribute to confidence are clearly defined.

There are currently two main frameworks used for the study of confidence, one based on Signal Detection Theory (SDT), and the other based on evidence accumulation (for a review, see Mamassian, 2016). The SDT framework (Green & Swets, 1966) has been exceedingly successful

for modelling choice tasks, also referred to as Type 1 tasks, and it also formed the basis for discussing confidence judgments, also known as Type 2 judgments (Clarke et al., 1959; Galvin et al., 2003). However, this framework is not intended to be a process model that describes *how* Type 2 judgments are actually made. The primary aim of the present manuscript is to provide a complete generative model for perceptual confidence judgments that is grounded in SDT. With this generative model, we have three main objectives that we briefly introduce next. These objectives are respectively: the separation of primary and secondary information for the computation of confidence, the construction of a measure of confidence efficiency that is defined at the metacognitive level, and the estimation of one critical form of confidence bias.

Our model of confidence is based on the idea that confidence judgments are derived from the current perceptual decision and some decision variable that we have called confidence evidence. The question of how this confidence evidence is formed has been central in prior research. In particular, confidence evidence is not necessarily identical to the sensory evidence used to make the Type 1 decision. This idea was supported by dissociations between confidence and decision accuracy documented in many studies. For instance, confidence judgments might fail to incorporate adequately the variance of evidence (e.g. de Gardelle & Mamassian, 2015; Spence, Dux, & Arnold, 2016; Boldt, de Gardelle, & Yeung, 2017), ignore sensory evidence going against the choice (e.g. Zylberberg, Barttfeld, & Sigman, 2012; Maniscalco, Peters, & Lau, 2016), and fail to properly track the fluctuations of attention (e.g. Wilimzig et al., 2008; Zizlsperger, Sauvigny, & Haarmeier, 2012; Recht, Mamassian, & de Gardelle, 2019). Confidence can also be influenced by information occurring after the Type 1 decision has been made (Resulaj et al., 2009; Pleskac & Busemeyer, 2010; Moran, Teodorescu, & Usher, 2015). These findings contributed to the appreciation that confidence evidence should be differentiated from sensory evidence, and that confidence evidence might be influenced by information, relevant or not, that was not used during the Type 1 decision. We note that one prior study investigating visibility ratings in a meta-contrasting masking paradigm has discarded such a “dual-channel model” (Maniscalco & Lau, 2016). However, visibility and confidence judgments are not necessarily equivalent, theoretically or empirically (Rausch & Zehetleitner, 2016). Besides, studies about perceptual confidence often find participants for whom metacognition is better than what is prescribed by performance (see e.g. Palmer, David, & Fleming, 2014; Hainguerlot et al., 2018; Moreira et al., 2018; see also Scott et al., 2014 for a similar result in a non-perceptual task). These observations indicate that confidence incorporates sometimes more information than what is used during the perceptual decision, and call for an additional channel of information.

In our model, we thus separate two streams of information, with one stream corresponding to the information that contributed to the Type 1 decision and another stream that provides additional information. We call these two streams *primary* and *secondary*. Through the primary stream, confidence evidence is just a duplicate of the sensory evidence that is used for the perceptual

decision. This stream of processing is present is all models of confidence. In contrast, through the secondary stream, confidence evidence has access to additional information. This additional information can result from some parallel processing of the stimulus properties and sensory information accumulated after the perceptual decision. The combined information from these two streams is also affected by some confidence noise, and by biases, as detailed below. The first objective of our modelling effort is thus to clarify the respective contributions of the primary and secondary streams to confidence judgments, both theoretically and empirically.

Figure 1 illustrates our modelling approach and provides the links between the different variables of the model. All the notations of the model are summarized in Table 1. We highlight in particular two components in relation to the primary and secondary confidence evidence. The first component is the *confidence noise* which characterizes the inefficiency of the confidence evidence computation relative to the *ideal confidence observer*. The second component is the *confidence boost* which characterizes the relative contribution of the secondary confidence evidence as a fraction of the overall confidence evidence. The reason why this latter component is called confidence boost is because new evidence from the stimulus will augment the information present at the Type 2 level and boost metacognitive efficiency towards a *super-ideal* level.

Confidence boost and confidence noise have opposite effects on Type 2 performance, and therefore it is difficult to properly estimate both of them in practice. Yet, it is important to have at our disposal an overall measure of Type 2 efficiency. Defining such a measure has been challenging in the past (Fleming & Lau, 2014), but a significant step forward was obtained recently thanks to the meta-d' computation (Maniscalco & Lau, 2012). This methodological tool allows experimenters to measure metacognitive abilities without confounds from Type 1 performance. However, one key characteristic of this measure is that it uses the metric of the Type 1 task, rather than of the Type 2 task. The second objective of our modelling effort is thus to offer a measure of Type 2 efficiency that is really anchored to the Type 2 level of processing.

The third objective of our modelling effort is to be able to detect and quantify some confidence biases. In our model, we focus on one particular type of confidence biases, where an over-confidence represents an over-estimation of one's perceptual sensitivity, or equivalently an under-estimate of the sensory noise. This type of confidence biases is difficult to detect because all the confidence judgments for a particular task are affected. When confidence is compared across two distinct tasks, we can obtain an estimate of the over-confidence for one task relative to the other. This kind of confidence comparison forms the basis of the confidence forced-choice paradigm. In this procedure, participants complete two Type 1 decisions on distinct stimuli, and then indicate which decision was associated with the greater confidence (Barthelmé & Mamassian, 2009; de Gardelle & Mamassian, 2015). We apply our generative model to the confidence forced-choice

paradigm and discuss how reliably each parameter of the model can be estimated in this paradigm.

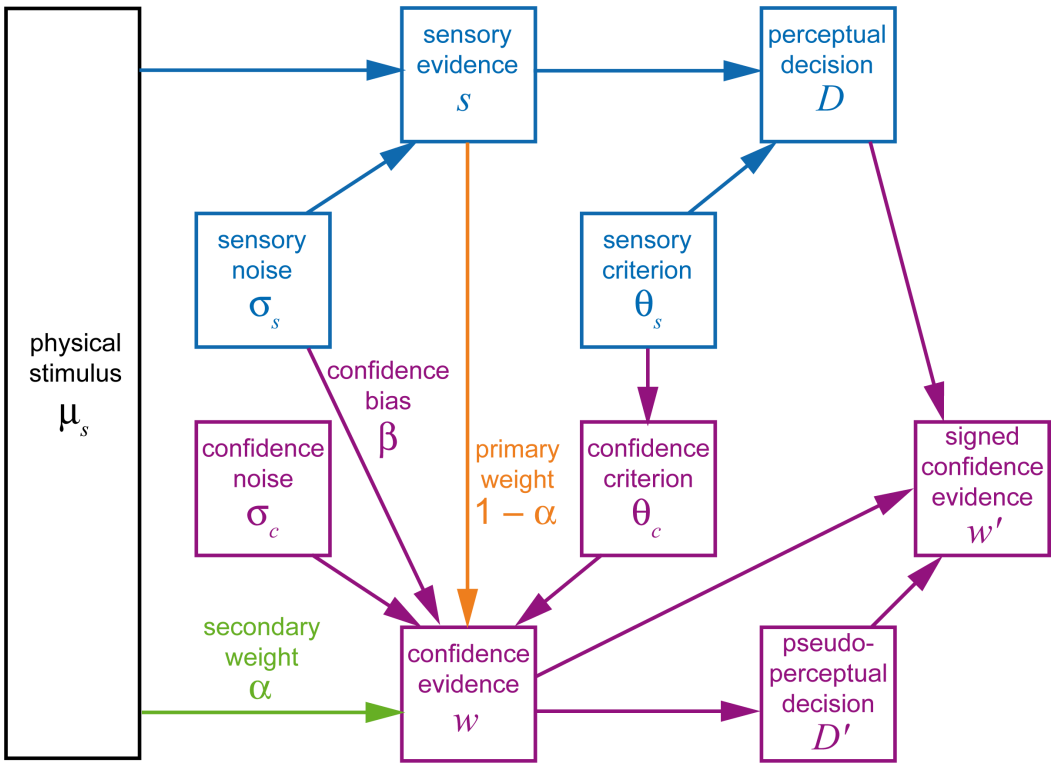

Figure 1. Overall framework for perceptual and confidence decision making. For Type 1 processing (in blue), the perceptual decision is based on sensory evidence that is an estimate of the physical stimulus. Sensory evidence is corrupted by sensory noise. For Type 2 processing (in purple), the confidence judgment is based on confidence evidence that is a combination of primary (orange) and secondary (green) evidence. The primary confidence evidence duplicates the sensory evidence whereas the secondary confidence evidence includes benefits from another look at the physical stimulus and other sources of information not illustrated here but described in the text. Confidence evidence is corrupted by additive confidence noise. It is also normalized by an estimate of sensory noise that is possibly corrupted by a multiplicative confidence bias, and it is compared to a confidence criterion that possibly differs from the sensory criterion. Finally, the signed confidence evidence is the magnitude of the confidence evidence that acquires a negative sign if the perceptual decision is incompatible with confidence evidence. See text for details.

As we compute confidence efficiency, we will see that the same confidence efficiency level can be achieved as a trade-off between confidence noise and confidence boost. The trade-off between these two parameters is a generic problem of confidence parameters indeterminacy.

Notation	Meaning	Domain
μ_{s}	Stimulus strength	$(-\infty, +\infty)$
s	Sensory evidence	$(-\infty, +\infty)$
w	Confidence evidence	$(-\infty, +\infty)$
w'	Signed confidence evidence	$(-\infty, +\infty)$
σ_{s}	Sensory noise (standard deviation of normal distribution) that drives perceptual sensitivity	$[0, +\infty)$
θ_{s}	Sensory criterion that drives bias in the perceptual decision	$(-\infty, +\infty)$
D	Perceptual decision based on sensory evidence	
D'	Pseudo perceptual decisions based on confidence evidence	
C	Confidence choice, i.e. interval chosen as more confident with respect to the self-consistency of the perceptual decision	$\{1, 2\}$
σ_{c}	Confidence noise (standard deviation)	$[0, +\infty)$
θ_{c}	Confidence criterion against which confidence evidence is evaluated	$(-\infty, +\infty)$
α	Confidence boost, i.e. the fraction of super-ideal confidence performance	$[0, 1]$
β	Confidence bias in over-estimating one's sensory sensitivity	$(0, +\infty)$
γ	Interval bias in favor of interval 1 in a confidence pair	$(-\infty, +\infty)$
$F(s_1,\ s_2)$	Joint distribution of sensory evidence in confidence pair	
$G(w_1, w_2 \mid s_1, s_2)$	Joint distribution of confidence evidence w conditional on sensory evidence s in confidence pair	
$H(s, w)$	Joint distribution of sensory and confidence evidence (its covariance matrix is K)	
$Q(s;\ \mu_s,\ \sigma_{\mathrm{s}})$	Mean of the distribution of confidence evidence conditional on a particular value of sensory evidence s	$(-\infty, +\infty)$
τ	Equivalent confidence noise (standard deviation)	$[0, +\infty)$
η	Confidence efficiency	$[0, +\infty)$

Table 1. Notations used in this manuscript.

Our manuscript is organized as follows. In the next two sections, we define what we mean by confidence in this manuscript, and then review briefly the confidence forced-choice paradigm. In section 4, we define the confidence ideal and super-ideal observers, which will help us determining the different ways confidence computation can be inefficient. We then detail our generative model in sections 5 and 6, describing how confidence evidence is linked to sensory evidence, and in sections 7 and 8, we apply this model to the confidence forced-choice paradigm. Section 9 explains the problem of confidence parameters indeterminacy and introduces how confidence efficiency is computed. We finish by showing the robustness of the parameter estimation (section 10), including the confidence bias (section 11), and illustrate in section 12 how the model can be fitted to real data by re-analysing one of our previous studies. Finally, section 13 presents a discussion of our approach.

2. Defining Confidence as Subjective Self-Consistency

We start by formally defining confidence in a perceptual decision as the subjective estimation made by an observer that her decision is self-consistent. Here, self-consistency refers to an agreement between the current perceptual decision and the most frequent decision made by the observer for a given stimulus and experimental conditions. Perceptual confidence is thus an estimation of the probability that the same decision would be made again, given the same physical stimulus and experimental conditions. In terms of Signal Detection Theory (SDT), self-consistency relates to perceptual sensitivity, disregarding perceptual bias.

Note that our definition slightly departs from the classic definition of confidence as an estimate of perceptual accuracy (i.e. probability of being correct). The difference between the two definitions is best illustrated by considering cases of perceptual illusions due to a sensory bias. In such cases, observers can be consistently incorrect in their decisions but still relatively confident in their perception. By focusing on self-consistency, rather than accuracy, our definition does not force us to call all observers overconfident in this case, which may be desirable given that the bias arises here at the perceptual level and not at the metacognitive level *per se*. If we follow the classic definition of confidence, however, we would have to conclude that the observer is overconfident because she is both incorrect and very confident.

Our definition of confidence as an estimate of one's own self-consistency aligns with other works. In meta-memory, Koriat (2012) has highlighted that confidence may reflect the consensuality of one's own answer with respect to answers chosen by other individuals, rather than just whether one's answer is correct or not. Our discussion of overconfidence is also reminiscent of one

particular type of overconfidence discussed in the literature. Three types of overconfidence are sometimes distinguished, namely the *overestimation* of one's accuracy, the *overplacement* relative to others, and the *overprecision* of one's beliefs (Moore & Healy, 2008). Our definition of confidence as subjective self-consistency naturally fits with overprecision. In other words, with our definition, an individual would be overconfident in a perceptual task if she overestimates her own sensitivity in this task. By contrast, the traditional definition of confidence as the subjective probability of being correct corresponds to overconfidence being an overestimation of one's accuracy. Note that in the SDT framework, these two definitions would be equivalent if all decision criteria are neutral. However, as detailed below, our modelling approach will allow for any criteria, including criteria that differ between Type 1 and Type 2 evaluation of the evidence.

As a final word, we note that if there is no sensory bias and if there is only one main interpretation for a stimulus, then self-consistency is the same as correctness. This could simplify the understanding of our model for readers who are not satisfied or confused with our definition of confidence.

3. Confidence Forced-Choice

In this manuscript, we focus on the confidence forced-choice paradigm. One key advantage of this procedure is to bypass the rating scale typically used to measure confidence, and to focus directly on the internal confidence, eliminating the need for participants to maintain a constant mapping between internal confidence and ratings. In this paradigm, participants indicate which of two intervals produces the highest feeling of confidence, where each interval consists of a stimulus, and a decision made on that stimulus. A *confidence trial* is thus composed of two stimuli, two perceptual decisions, and the confidence comparison choice between these two decisions. Therefore, this procedure requires participants to hold in working memory their confidence judgment in the first interval to be able to compare it to the one in the second interval. However, it is unclear whether this cost is higher than the one to set and hold in memory a set of stable confidence rating criteria.

Let us consider a typical use of the confidence forced-choice paradigm around a psychophysical experiment. In this example, the perceptual task is to indicate whether the dots of a random-dot kinematogram stimulus are moving to the right or to the left relative to a reference direction. We code rightward motions with positive values and leftward motions negatively. Across trials, stimuli differ in strength, i.e. in how much the motion direction deviates from the reference. Stimulus strength affects how well observers can discriminate the direction of motion, as represented by the psychometric function (Figure 2A). The slope of the psychometric function reflects the sensitivity of the observer in the perceptual task.

To examine how confidence relates to perceptual sensitivity, we can analyze separately the perceptual decisions associated with higher and lower confidence in each confidence trial. We can then replot the psychometric function separately for these *confidence-chosen* and for *confidence-declined* decisions (Figure 2B). In the example of the figure, these two new psychometric functions are distinct, the one for the confidence-chosen decisions presents a steeper slope than the one for the confidence-declined decisions, or than the original one estimated over all trials (Figure 2A). This property is a signature of meta-perception, as it indicates that participants were able to pick the interval that led to a better performance, at least for some trials. If the participants gave their metacognitive judgments at random, as if they were not able to judge the quality of their perceptual decisions, the psychometric functions for chosen and declined decisions would overlap completely. In contrast, when the observer is using all the information she can use for her confidence judgment, the gain in the slope of the psychometric functions is strictly larger than zero.

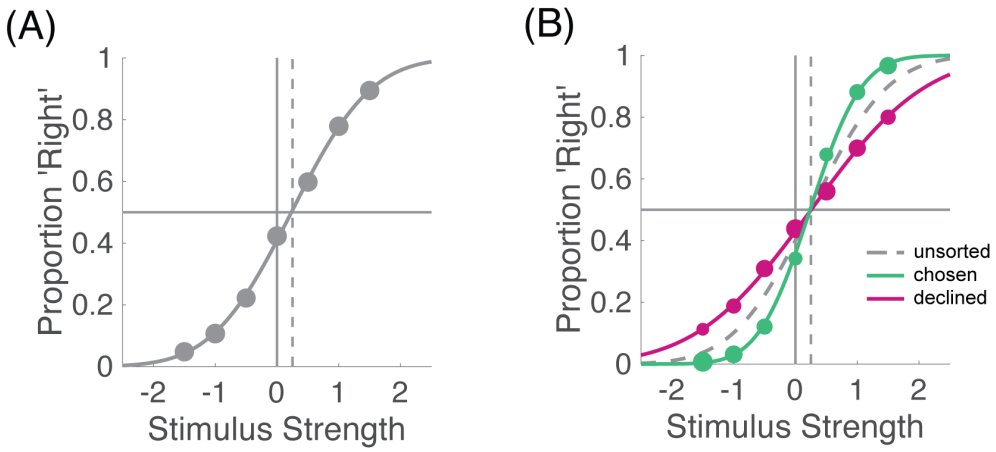

Figure 2. Psychometric functions. (A) Original psychometric function. The psychometric function links stimulus strength to perceptual decision, here the proportion of dots moving rightward. The solid line is a cumulative Gaussian fit to the psychometric functions. The standard deviation of the best fit determines its slope (here 1.01, a good approximation of the parameter $1/\sigma_s$ used in the simulation). (B) Psychometric function split by confidence. Trials judged to have higher confidence are sorted out and a new psychometric function is plotted for these trials only (green points). The remaining trials have been declined for confidence (red points). Dots size is proportional to the number of trials in each condition. For the psychometric function based on the chosen trials for confidence, the best fit gives a slope of 1.47. The gain in the slope of the psychometric functions from the unsorted (grey dashed curve) to the chosen (green curve) trials is therefore $1.47/1.01 = 1.45$. The parameters used to generate this and the following figures are provided in Table 2.

Even though it is simple and natural to use the gain in the slope of psychometric functions as an index of metacognitive ability (see, e.g. Barthelmé & Mamassian, 2009; De Martino et al., 2013; de Gardelle & Mamassian, 2014, 2015), we introduce later the confidence efficiency as an alternative descriptor of confidence sensitivity. Indeed, the comparison of psychometric functions actually discards important information about which confidence pairs were presented to participants. The full data set includes not only how a given perceptual trial falls into the confidence-chosen or confidence-declined set, but also how the confidence comparison choice depends on the two trials within a pair, which may have different stimulus strengths and different decisions. In the example of the simulated experiment shown in Figure 2, there were 7 possible stimulus strengths and two possible perceptual decisions ('R' or 'L') for each interval in a confidence pair, leading to 196 (7 x 7 x 2 x 2) possible combinations. In each of these combinations, we can measure the probability that interval 1 is associated with a greater confidence than interval 2.

All the confidence choice probabilities for the experiment summarized in Figure 2 are illustrated in Figure 3. This figure has four separate panels, one for each combination of the perceptual decisions in the first and second intervals (for instance, the top-left panel corresponds to the case where the observer saw leftward motion in the first interval and rightward in the second). The x-axis represents the stimulus strength in the first interval, and the curves of different colors correspond to different stimulus strength in the second interval. The y-axis shows the probability of choosing the perceptual decision in interval 1 as more confident than the one in interval 2. In the top-left panel, because the observer's decision was leftward in interval 1, this decision is more likely to be correct when the stimulus strength in interval 1 is more negative, which is why the probability of choosing interval 1 rises from right to left. Similarly, in that same panel, the perceptual decision was rightward in interval 2, so it is more likely to be correct when the stimulus strength in interval 2 is more positive, which is why the probability of choosing interval 1 declines from the top to the bottom curve.

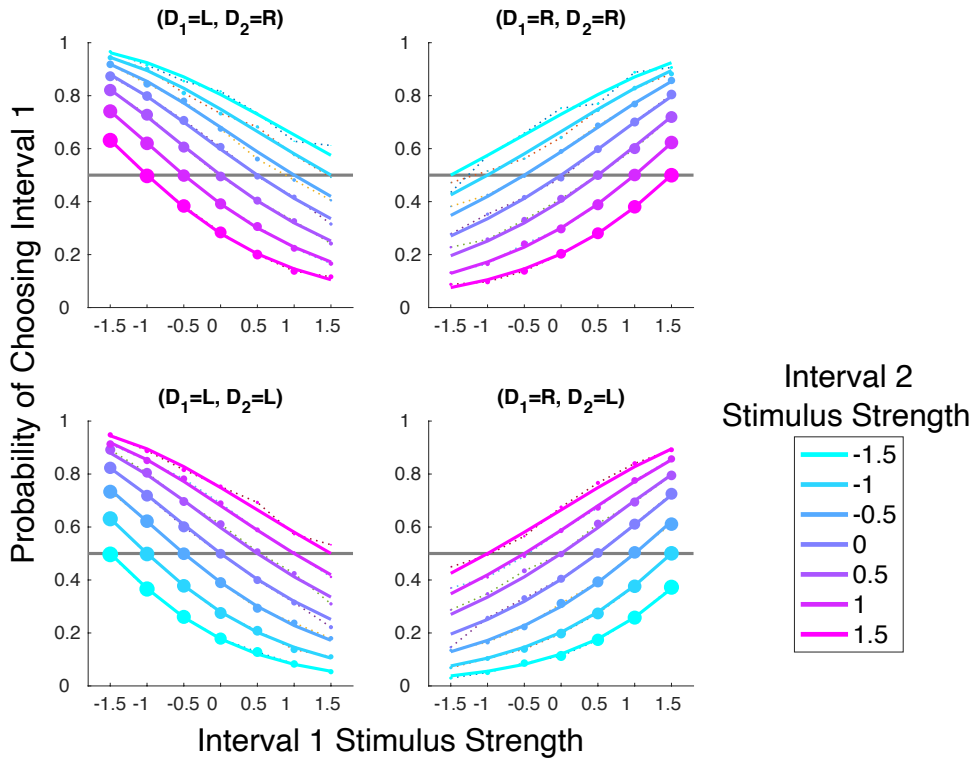

Figure 3. Confidence choice probabilities for each combination of stimulus strengths. Each panel shows the probability of choosing the first interval as the more confident one, given the stimulus strength presented in the first interval (x-axis) and in the second interval (colored lines). The four panels correspond to the different pairs of perceptual decisions across the two intervals (e.g. responses $D_1 = \text{'L'}$ and $D_2 = \text{'R'}$ in the top left panel). Dot size is proportional to the number of trials obtained in the simulation for this particular combination of stimulus strength and perceptual decision. Dotted lines link points that have the same stimulus strength in the second interval. The solid curves show the best fitted model described later in the manuscript. In this plot, parameters are those listed in Table 2, except $n = 100{,}000$ for figure clarity.

From the simulations shown in Figure 3, we see that confidence depends on the interaction between stimulus strength and perceptual decision, as typically found in empirical data. To better illustrate this pattern, let us focus on one subset where the stimulus strength in the second interval is 0 and the perceptual decision for this stimulus is 'R'. This subset corresponds to the middle blue line in the two top panels, which are replotted on Figure 4 but in different colors. Specifically, self-consistent perceptual decisions are shown in green, and self-inconsistent decisions in red. Here, self-consistent decisions in interval 1 correspond to responding 'R' for stimulus strengths in the first interval that are above the sensory criterion (0.25), and responding 'L' below. As expected, the

probability of choosing the first interval with greater confidence is always larger for self-consistent than for self-inconsistent decisions. In addition, as stimulus strength deviates more from the sensory criterion, confidence increases for self-consistent decisions, and decreases for self-inconsistent decisions. This is expected from a participant who displays meta-perception, although the exact form of this X-pattern varies across experimental conditions and models of confidence (Sanders et al., 2016; Adler & Ma, 2018; Rausch & Zehetleitner, 2019).

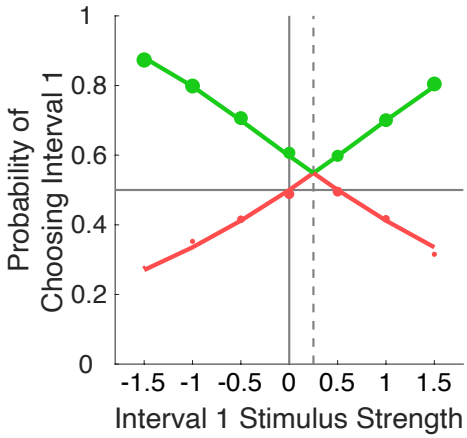

Figure 4. Choice probabilities for self-consistent and inconsistent decisions. The plot shows the same data as in Figure 3, for one particular sensory stimulus and perceptual decision in interval 2. Self-consistent perceptual decisions in interval 1 are shown in green, and self-inconsistent decisions in red. The probability of choosing the first interval with greater confidence is larger for self-consistent than for self-inconsistent decisions. In addition, as stimulus strength deviates more from the sensory criterion, confidence increases for self-consistent decisions, and decreases for self-inconsistent decisions. The resulting X-pattern is symmetric about the vertical axis passing through the sensory criterion.

We are now interested in modelling the sensitivity with which participants can estimate their confidence in their perceptual decisions. The model will attempt to replicate all 196 different probabilities that interval 1 is the winner of the confidence decision in Figure 3. In particular, we are interested in describing the ideal confidence observer that is using the exact same information for confidence judgments as the perceptual decisions, so that we can compare human meta-perceptual sensitivity to this ideal confidence observer. Along the way, we will also define a super-ideal confidence observer that maximizes confidence performance. Unless otherwise noted, the parameters in the figures take the default values shown in Table 2.

Parameter	Meaning	Figure Value	Ideal Value
$\{\mu_{\mathrm{A}},\ \mu_{\mathrm{B}}\}$	Examples of stimulus strengths	$\{1.5, -0.5\}$	
μ_{s}	Stimulus strengths for a complete simulated experiment	$-1.5\colon 0.5\colon 1.5$	
$(s_1,\ s_2)$	Sensory evidence in intervals 1 and 2 of a confidence pair where stimulus strengths are $(\mu_{\mathrm{A}},\ \mu_{\mathrm{B}})$	$(0.9, 0.7)$	
(D_1, D_2)	Perceptual decisions in intervals 1 and 2 of a confidence pair where stimulus strengths are $(\mu_{\mathrm{A}},\ \mu_{\mathrm{B}})$	(R, R)	(R, L)
σ_{s}	Sensory noise (standard deviation)	1.0	0.0
θ_{s}	Sensory criterion that drives bias in the perceptual decision	0.25	0.0
σ_{c}	Confidence noise (standard deviation)	0.5	0.0
θ_{c}	Confidence criterion	0.0	0.0
α	Confidence boost	0.2	0.0
β	Confidence bias in over-estimating one's sensory sensitivity	1.0	1.0
γ	Interval bias in favor of interval 1 in a confidence pair	0.0	0.0
n	Number of confidence pairs in a simulation	$10{,}000$	

Table 2. Unless explicitly stated in the figure caption, the parameter values used in the figures are the ones in this table. In the last column are shown the values corresponding to the ideal observer and ideal confidence observer.

4. Ideal Confidence Observer

In this section, we present how the perceptual decision is derived from sensory evidence. By analogy, we introduce the confidence evidence that will be the basis for the confidence judgment. The approach is based on Signal Detection Theory (Green and Swets, 1966) and ideal observer principles (Barlow, 1962; Geisler, 1989). In the next section, we will generalize this model of confidence by considering several ways in which actual confidence judgments can deviate from optimal ones.

4.a. Perceptual Decisions

We consider here a perceptual task in which a stimulus has to be categorized as 'Right' ('R') or 'Left' ('L'). In a typical psychophysical experiment, there will be a range of stimuli with different levels of difficulty that we represent by the stimulus strength μ_s. For instance, stimuli could be random dot kinematograms where each dot has a motion direction drawn from a circular normal distribution whose mean μ_s is slightly clockwise (or 'Right') or counter-clockwise ('Left') of the vertical upward direction. For illustrative purposes in this part of the report, we first consider two such stimuli, A and B, that belong to categories 'R' and 'L' respectively (Figure 5).

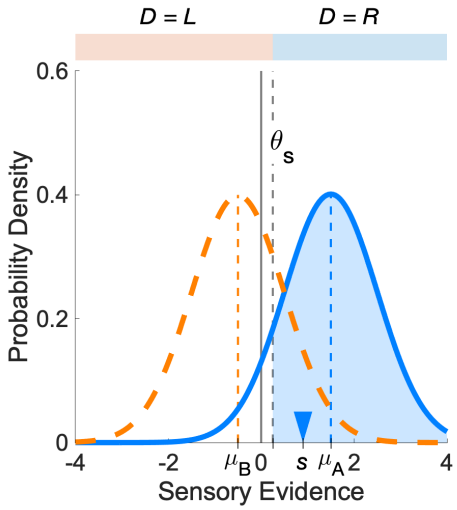

Figure 5. Sensory evidence in a perceptual discrimination task. Stimuli to be discriminated belong to two categories 'Right' ('R') and 'Left' ('L'). The distribution of sensory evidence for two stimuli A and B is in blue and orange, respectively. On each trial, the participant has access to one sample of the stimulus category presented on that trial (a sample s from stimulus A is shown by the blue triangle). All sensory evidence to the right of sensory criterion θ_s, represented by the blue shaded area, are assigned to the 'R' category.

Because of sensory noise, the observer only has access to some noisy sensory evidence s. We assume that on average the observer has an unbiased estimate of the sensory strength, so the mean of s is μ_s. For simplicity, we further assume that the sensory noise is normally distributed, with common variance σ_s^2 for all stimuli, such that a sensory noise sample ϵ_s for one particular trial follows the distribution $\epsilon_s \sim \mathrm{N}(0,\ \sigma_s^2)$.

The sensory evidence on one trial is then

$$s = \mu_s + \epsilon_s \ , \tag{1}$$

where $\mu_s = \mu_A$ if stimulus A was presented, and $\mu_s = \mu_B$ if stimulus B was presented instead. A perceptual decision (Type 1 decision D) consists in comparing the sensory evidence against a sensory criterion θ_s, namely

$$\begin{cases} D = \text{'R'} \quad \text{if } s > \theta_s \ , \\ D = \text{'L'} \quad \text{otherwise} \end{cases} . \tag{2}$$

The most frequent percept for stimulus A is 'R' (the blue shaded area in Figure 5 to the right of the sensory criterion is larger than 0.5 because $\mu_A > \theta_s$). Therefore, when stimulus A is presented, the perceptual decision will be self-consistent if it is 'R'. We present other properties of self-consistency in Appendix A.

4.b. Ideal Confidence Observer

Now that we have modelled perceptual decisions, we can consider confidence (Type 2) judgments. We start with the important case of the *ideal confidence observer* that will be used as a reference to compare human confidence judgments. The ideal confidence observer is ideal for its confidence judgment but suboptimal for its perceptual decision. In other words, this particular observer has the same sensory sensitivity and biases as the human observer, and thus is similarly subject to sensory noise and sensory criterion shifts as the human observer. However, it is ideal in the sense that it is able to judge optimally which of two perceptual decisions is more likely to be self-consistent based on the same sensory information that has been used to reach the perceptual decisions. In other words, for the ideal confidence observer, the confidence evidence will be entirely determined by the sensory evidence.

From Figure 5, we see that the perceptual decision is more likely to be self-consistent when the sensory evidence s is further away from the sensory criterion θ_s (for a formal description of the probability of being self-consistent, see Appendix A). Therefore, from the point of view of the ideal confidence observer, the distance of the sensory evidence to the perceptual decision boundary is a good decision variable to estimate confidence (Galvin et al., 2003). We follow this tradition with one particular twist. To be able to estimate confidence sensitivity irrespective of the sensory sensitivity of the observer for the current task, we normalize the distance to the decision boundary by the sensory noise. As can be seen in Appendix D, this step alleviates apparent contradictions such that sensory noise increases metacognitive efficiency (Bang et al., 2019). In summary, we define the ideal confidence evidence to be

$$w_{\text{ideal}} = (s - \theta_s)/\sigma_s \ . \tag{3}$$

Because confidence evidence has been normalized by sensory noise, it is a unit-free measure of confidence. In other words, it is not bound to the stimulus dimension that is relevant for a task (e.g. the angle in degrees of motion direction if the task of the observer is to estimate motion direction). This property is useful when comparing confidence across tasks (de Gardelle & Mamassian, 2014). Further motivation for this choice of ideal confidence evidence is presented in Appendix A.

4.c. Super-Ideal Confidence Observer

In contrast to the ideal confidence observer, the *super-ideal confidence observer* has access to the original stimulus, and not just the noisy sensory evidence used to make the perceptual decision. This scenario can actually lead to better performance than the ideal confidence observer, hence the term "super-ideal" confidence observer. This extreme scenario is interesting to consider because confidence judgments are often performed after perceptual decisions, and thus can benefit from a more extensive analysis (e.g. Pleskac & Busemeyer, 2010) or second look at the stimulus. Confidence evidence for the super-ideal confidence observer is now

$$w_{\text{super\_ideal}} = (\mu_s - \theta_s)/\sigma_s \quad . \tag{4}$$

In this definition, the super-ideal confidence observer is not corrupted by any sensory noise. While it is implausible that any human observer will ever be able to access the stimulus without some form of noise, it is still important to set such an upper-bound on confidence performance. We should also remark that any sensory noise that may corrupt the super-ideal confidence observer will be absorbed in the confidence noise that we will define later (see Appendix F and below for the definition of confidence noise).

Note that we still normalize the stimulus strength μ_s relative to the sensory noise σ_s and sensory criterion θ_s so as to obtain a unit-free measure of confidence that still reflects the potential perceptual bias of the observer.

5. Generative Model of Confidence Evidence

In the previous section, we have described the ideal and super-ideal confidence observers. We now consider four ways in which human confidence judgments can deviate from the ideal confidence observer. First, human observers can behave partially as the super-ideal confidence observer, thereby boosting their confidence sensitivity. Second, they can display some confidence noise that is impairing their ability to use their confidence evidence. Third, human observers can be inaccurate in their estimate of the sensory sensitivity, thereby generating over- or under-

confidence. Finally, human observers can be inaccurate in their estimate of the sensory bias, thereby creating potential conflicts between sensory and confidence decisions. We now examine these four cases in turn.

5.a. Confidence Boost

We define *confidence boost*, noted α, the fraction of the super-ideal confidence observer that contributes to the human confidence evidence. If $\alpha = 1$, then the human observer is just like the super-ideal confidence observer, and if $\alpha = 0$, then the human observer behaves just like the ideal confidence observer. Confidence evidence now becomes a mixture of the evidence from the super-ideal and ideal confidence observers, namely

$$w = \alpha \,.\, w_{\text{super\_ideal}} + (1 - \alpha) \,.\, w_{\text{ideal}} \quad . \tag{5}$$

Before we proceed further, we should clarify that this formulation of a weighted sum of the super-ideal and ideal components is mainly a convenient mathematical way to describe all the information that can contribute to the confidence evidence. It is not, however, an assumption about the actual psychological mechanism underlying confidence judgments. In particular, we do not need to assume that observers have a separate and direct access to the stimulus when taking a second look at the stimulus. More likely, observers would refine their estimate of the validity of their perceptual decision by combining multiple pieces of noisy confidence evidence. In Appendix F, we present a more detailed generative model that distinguishes different sources of confidence evidence.

The expression of the confidence evidence in Equation 5 can be rewritten as

$$w = (\alpha \,.\, \mu_s + (1 - \alpha) \,.\, s - \theta_s)/\sigma_s$$

$$w = (\mu_s + (1 - \alpha) \,.\, \epsilon_s - \theta_s)/\sigma_s \quad . \tag{6}$$

The effect of confidence boost on the psychometric function is shown in Figure 6A. This psychometric function should be compared to the one with the default parameters in Figure 2B. When confidence boost increases, we observe a steeper psychometric function for the confidence-chosen trials. In other words, the observer is better able to discriminate correct from incorrect perceptual decisions. This is not surprising as the confidence boost reflects the ability of the observer to use more information at the metacognitive level.

5.b. Confidence Noise

Just like sensory noise corrupts the sensory evidence, we introduce *confidence noise* that corrupts the confidence evidence. We model confidence noise as a zero-mean normal distribution with variance σ_c^2, such that a confidence noise sample ϵ_c follows the distribution $\epsilon_c \sim N(0, \sigma_c^2)$. We assume that confidence noise is additive and independent of sensory evidence, so the new confidence evidence becomes

$$w = (\mu_s + (1-\alpha).\epsilon_s - \theta_s)/\sigma_s + \epsilon_c \ . \quad (7)$$

Because confidence noise is unrelated to the sensory evidence, it is unit-less, and comparable across different tasks (see e.g. de Gardelle & Mamassian, 2014). The effect of confidence noise on the psychometric function is shown in Figure 6B. When the confidence noise increases, we obtain a shallower psychometric function for the confidence chosen trials. In other words, the observer is less able to discriminate correct from incorrect perceptual decisions.

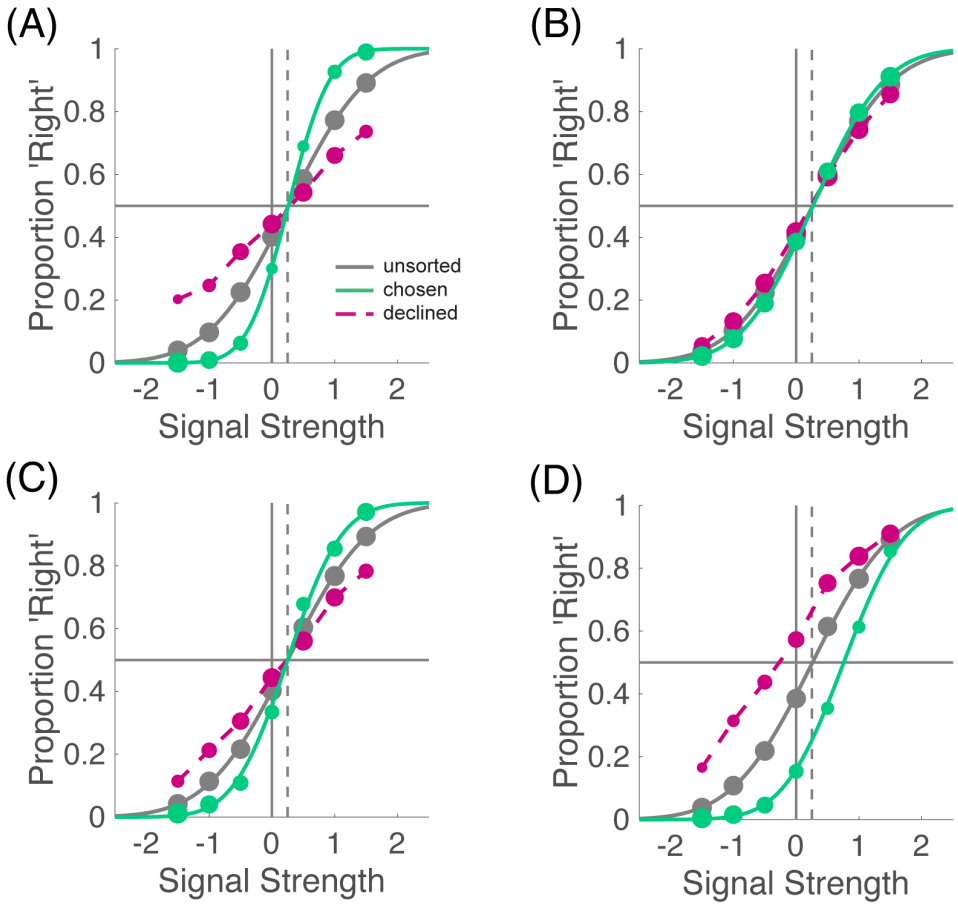

Figure 6. Influence of different model parameters on the psychometric functions. In these plots, the parameters are those listed in Table 2, except for one parameter.

Plotting conventions are the same as those used in Figure 2B. (A) The confidence boost is increased to $\alpha = 0.8$. (B) The confidence noise is increased to $\sigma_c = 2.0$. (C). The confidence bias is increased to $\beta = 2.0$. (D). The confidence criterion is increased to $\theta_c = 1.0$.

5.c. Confidence Bias

Sensory evidence needs to be scaled to generate the confidence evidence such that the latter is task-independent and unit-free. This is achieved by normalizing confidence evidence relative to the sensory sensitivity, and consequently, confidence evidence is a good proxy for the probability of being self-consistent in the perceptual decision (see again Appendix A). From the ideal confidence observer perspective, this scaling factor should be the inverse of the sensory noise ($1/\sigma_s$). We represent by β the *confidence bias* which stands as a deviation away from this ideal scaling (this corresponds to replacing $1/\sigma_s$ with β/σ_s). Values of β larger than 1.0 indicate over-confidence, and values smaller than 1.0 under-confidence. Considering this misestimate of the sensory sensitivity leads to a new confidence evidence

$$w = (\mu_s + (1-\alpha).\epsilon_s - \theta_s).\beta/\sigma_s + \epsilon_c \quad . \tag{8}$$

The effect of confidence bias on the psychometric function is shown in Figure 6C. We observe that the psychometric function for the confidence chosen trials is not affected by the confidence bias (Figure 6C is identical to Figure 2B). This is not surprising since this parameter scales the confidence evidence in both intervals in the same way. Even though the effects of confidence bias are invisible here, we present below a condition where this confidence bias can be partially estimated (see section 11).

5.d. Confidence Criterion

Finally, human observers can use a criterion against which they measure their confidence that is distinct from the sensory criterion. We represent by θ_c the deviation of the *confidence criterion* away from the sensory criterion. Ideally this parameter is zero ($\theta_c = 0$), but when it is not, the confidence evidence becomes

$$w = (\mu_s + (1-\alpha).\epsilon_s - \theta_s - \theta_c).\beta/\sigma_s + \epsilon_c \quad . \tag{9}$$

The effect of confidence criterion on the psychometric function is shown in Figure 6D. When the confidence criterion deviates from the sensory criterion, the point of subjective equality (PSE) for the confidence-chosen decisions (green curve) becomes different from the PSE for the original psychometric function (grey curve). The shift in PSE is coming from the inconsistency between the perceptual decision and what we will call the "pseudo perceptual decision" (see section 6.c), for a range of sensory values near the sensory criterion.

6. Covariation of Sensory and Confidence Evidence

Because of noise at the perceptual level or at the confidence level, sensory evidence and confidence evidence will vary across trials, even when the stimuli and the responses are the same. We will now characterize this variation, by defining the joint distribution of sensory and confidence evidence. This will allow us to produce summary statistics that will be useful for presenting the full model of the confidence comparison task. We note that previous models of confidence have discussed the joint distribution between sensory and confidence evidence (Fleming & Daw, 2017). However, it is important to appreciate that our definition is different from these previous studies because our joint distribution is derived from a generative model based on the introduction of confidence noise and confidence boost instead of being an arbitrary bivariate distribution function.

6.a. Joint distribution for sensory and confidence evidence

Considering all the possible deviations from the ideal confidence observer, the confidence evidence is following Equation 9 above. This evidence is normally distributed with mean

$$\mathrm{E}[w] = (\mu_s - \theta_\mathrm{s} - \theta_\mathrm{c})\,\beta/\sigma_\mathrm{s}\ . \tag{10}$$

In addition, we note that confidence noise is independent of sensory evidence. This allows us to characterize the variance of the distribution of confidence evidence as

$$\mathrm{var}[w] = (1-\alpha)^2\beta^2 + \sigma_\mathrm{c}^2\ . \tag{11}$$

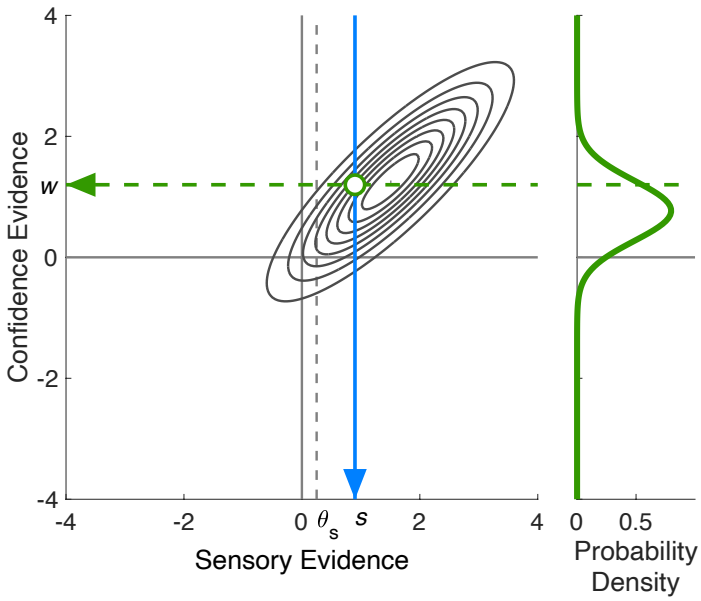

Figure 7. Joint distribution of sensory and confidence evidence. On each trial, the participant has access to one sensory sample s (blue arrow) and one confidence sample w (green arrow) of the joint distribution $H(s, w)$. The blue distribution shown in Figure 5 is the marginal distribution of the sensory evidence. The green distribution in the right-hand panel is the distribution of confidence evidence for the particular sensory sample $s = 0.9$ (it is the cross-section of the joint distribution along the blue line). The mean of this distribution is $Q(s;\, \mu_{\mathrm{A}},\, \sigma_{\mathrm{s}})$ (see below, Equation 17), and its spread is the confidence noise σ_c. The strength of the confidence evidence on that particular trial is given by the magnitude of the sample w (distance away from zero).

Because both the sensory and confidence evidence are normally distributed, their joint distribution $H(s, w)$ is a bivariate normal distribution. An example of this joint distribution is shown in Figure 7.

The mean of the joint distribution $H(s, w)$ is obtained from the mean of the sensory evidence and the mean of the confidence evidence (see Equation 10)

$$\mathrm{E}[\, H(s, w)\,] = [\mu_{\mathrm{s}}\, ,\ (\mu_{\mathrm{s}} - \theta_{\mathrm{s}} - \theta_{\mathrm{c}})\, \beta / \sigma_{\mathrm{s}}] \quad . \tag{12}$$

The covariance between s and w is obtained from Equation 9

$$\mathrm{cov}(s, w) = \ (1 - \ \alpha)\, \beta\, \sigma_{\mathrm{s}} \quad , \tag{13}$$

so that the covariance matrix K of the joint distribution H is

$$K = \mathrm{cov}[\, H(s,w)\,] = \begin{bmatrix} \sigma_s^2 & (1-\alpha)\,\beta\,\sigma_s \\ (1-\alpha)\,\beta\,\sigma_s & (1-\alpha)^2\beta^2 + \sigma_c^2 \end{bmatrix} . \quad (14)$$

It is worth noting the special case of the ideal confidence observer. In this case, $\alpha = 0$, $\beta = 1$, $\sigma_c = 0$, and the covariance matrix reduces to

$$K_{\text{ideal}} = \begin{bmatrix} \sigma_s^2 & \sigma_s \\ \sigma_s & 1 \end{bmatrix} . \quad (15)$$

The determinant of this covariance matrix is zero, indicating that there is a direct mapping between sensory evidence and confidence evidence: this is expected since without confidence noise, confidence and sensory evidence are perfectly correlated.

One other special case of interest is the super-ideal confidence observer ($\alpha = 1$) corrupted with some confidence noise, where the covariance matrix is

$$K_{\text{noisy\_super\_ideal}} = \begin{bmatrix} \sigma_s^2 & 0 \\ 0 & \sigma_c^2 \end{bmatrix} . \quad (16)$$

This covariance matrix is now diagonal, indicating that confidence and sensory evidences are independent. Here, the joint distribution H has its main axes oriented along the sensory and confidence evidence axes. In other words, for a noisy super-ideal confidence observer, confidence evidence depends only on the stimulus strength and is independent from the sensory evidence for the current trial.

6.b. Confidence Evidence Conditional on Sensory Evidence

On any perceptual trial of a confidence pair, the observer first gets some sensory evidence, performs a perceptual decision based on this sensory evidence, and then estimates the confidence that this decision is self-consistent. Therefore, we need to estimate the distribution of confidence evidence for one particular value of sensory evidence s (Figure 7, right-hand panel). This distribution of confidence evidence is $\mathrm{P}(w \mid s)$ and corresponds to a section of the joint distribution $H(s,w)$. This conditional distribution is normally distributed, and its mean (that we denote $Q(s;\, \mu_s,\, \sigma_s)$ for later use) and variance can be inferred from the mean and the covariance matrix of the joint distribution H (Equations 12 and 14)

$$\begin{cases} \mathrm{E}[w \mid s] = Q(s;\, \mu_s,\, \sigma_s) = [s + (\mu_s - s)\,\alpha - \theta_s - \theta_c]\;\beta\,/\,\sigma_s \\ \mathrm{var}[w \mid s] = \sigma_c^2 \end{cases} . \quad (17)$$

As expected, we see that the variance of the confidence evidence, once the sensory evidence is known, is just the variance of the confidence noise. The mean is a biased and scaled version of the sensory evidence s. It is biased towards the representation of the original stimulus μ_s when the parameter α is larger than zero, i.e. when the human confidence observer is behaving a bit like the super-ideal confidence observer. The scaling involves the parameter β that is responsible for a proper calibration of confidence judgments, such that $\beta > 1$ corresponds to over-confidence.

6.c. Pseudo-Perceptual Decision

The confidence evidence is the basis to judge whether the perceptual decision is self-consistent. One might be tempted to just use the absolute value of confidence evidence for this judgment, where larger absolute values reflect better chances to be self-consistent. However, this choice would disregard the actual perceptual decisions that were taken. Critically, to decide whether the perceptual decision is self-consistent, we need to evaluate whether the confidence evidence is consistent with the perceptual decision. For this purpose, we introduce the *pseudo perceptual decision* D' that corresponds to the perceptual decision that would have been taken if the confidence evidence was used instead of the sensory evidence. By similarity to the definition of perceptual decisions in Equation 2 above, the pseudo perceptual decision is thus defined as

$$\begin{cases} D' = \mathrm{R} & \text{if } w > 0\,, \\ D' = \mathrm{L} & \text{otherwise} \end{cases}\,. \tag{18}$$

When the pseudo perceptual decision D' is distinct from the perceptual decision D, this can be taken as an alert signal that the perceptual decision might be invalid. Therefore, we can define a new variable that reflects the diminished trust that the perceptual decision was valid when D' is distinct from D. We define the *signed confidence evidence* as

$$w' = \begin{cases} |w| & \text{if } D = D'\,, \\ -|w| & \text{otherwise} \end{cases}\,. \tag{19}$$

This signed confidence evidence is useful to estimate the probability that the perceptual decision is self-consistent given the current confidence evidence and perceptual decision, $\mathrm{P}(\text{self-consistent} \mid w, D)$. Computing this probability is complex because it rests on the knowledge of all the parameters in our model. Whereas prior work has assumed that observers would be able to use this knowledge (Fleming & Daw, 2017), here we propose instead that human observers only have access to the current level of confidence evidence and what they decided perceptually. Therefore, we propose that the observer is computing the *confidence probability* defined as

$$\mathrm{P}(\text{confident} \mid w, D) = \Phi(\mathrm{w}') \quad , \tag{20}$$

where Φ is the cumulative of the standard normal distribution. In Appendix A, we show that the confidence probability is a reasonable proxy for the probability of being self-consistent given the current confidence evidence and perceptual decision.

7. Comparing Confidence Across Two Perceptual Decisions

In the confidence forced-choice paradigm, two intervals are presented to the observer who has to choose the one for which she feels more confident that her perceptual decision was self-consistent. Therefore, we need to compare confidence across the two perceptual decisions of a confidence pair.

7.a. Joint Sensory Evidence and Joint Confidence Evidence in a Confidence Pair

Typically, the stimuli presented in the two intervals are independent from each other, so that we can assume that the sensory evidence in the two intervals is uncorrelated. Likewise, we assume that the confidence evidence in the two intervals is also uncorrelated[1]. It is convenient to represent sensory and confidence evidence across the two intervals as joint probability distributions (Figure 8).

[1] At this stage, this is a simplifying assumption that might be invalid (for some evidence, see e.g. Rahnev et al., 2015; Bosch et al. 2020). However, let us consider briefly two scenarios. First, there might be some positive correlation coming from slow fluctuations of confidence (due for instance to fluctuations of arousal). We do not regard these fluctuations as a critical problem because the two intervals will be similarly affected. Second, there might be some negative correlation due to an exaggerated attention to one of the two intervals. In our model, we have a parameter that reflects part of this issue, the interval bias parameter γ (see Appendix B and below for the definition of this parameter).

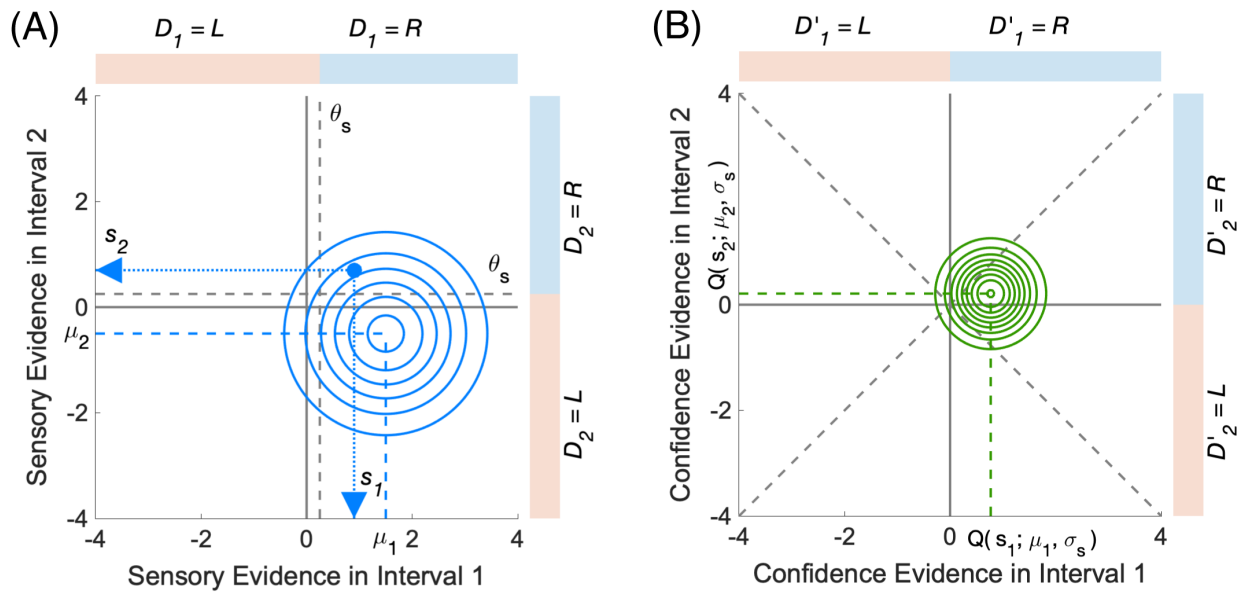

Figure 8. Joint distributions of sensory and confidence evidence across the two intervals of a confidence pair. (A) Joint distribution for the sensory evidence $F(s_1, s_2)$. In this example, stimulus A is presented in interval 1 ($\mu_1 = \mu_A$) and stimulus B is presented in interval 2 ($\mu_2 = \mu_B$), and are associated with the same level of sensory noise ($\sigma_1^2 = \sigma_2^2 = \sigma_s^2$). The joint distribution of the sensory evidence is shown as a contour plot in blue. A sample of this joint distribution is shown as a blue dot that has coordinates s_1 for interval 1 and s_2 for interval 2. The perceptual decisions D_1 and D_2 associated with this sample are both in favor of response R. (B) Joint distribution for the confidence evidence conditional on sensory evidence $G(w_1, w_2 \mid s_1, s_2)$. Because the perceptual decisions were R for both intervals, the joint confidence distribution is likely to have its centre in the upper-right quadrant (contour plot in green). The pseudo perceptual decisions D_1' and D_2' are shown for the confidence evidence space.

The joint distribution $F(s_1, s_2)$ for the sensory evidence across the two intervals is a bivariate normal distribution (Figure 8A) with mean and covariance

$$\begin{cases} \mathrm{E}[\, F(s_1, s_2)\,] = [\mu_1,\ \mu_2] \\ \mathrm{cov}[\, F(s_1, s_2)\,] = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \end{cases} . \tag{21}$$

The joint distribution $G(w_1, w_2 \mid s_1, s_2)$ is the confidence evidence conditional on the sensory evidence across the two intervals (Figure 8B). It is a bivariate normal distribution with mean and covariance matrix

$$\begin{cases} \mathrm{E}[\, G(w_1, w_2 \mid s_1, s_2)\,] = [\, Q(s_1;\, \mu_1,\, \sigma_1),\; Q(s_2;\, \mu_2,\, \sigma_2)\,] \\ \mathrm{cov}[\, G(w_1, w_2 \mid s_1, s_2)\,] = \begin{bmatrix} \sigma_c^2 & 0 \\ 0 & \sigma_c^2 \end{bmatrix} \end{cases} . \quad (22)$$

where the off-diagonal elements of the covariance matrix are zero because confidence evidence was assumed to be uncorrelated across intervals. The mean values are computed from Equation 17.

7.b. Confidence Decision Rule

The final step in choosing the interval in the confidence forced-choice paradigm is to decide on a *confidence decision rule*. This decision rule uses the confidence evidence in both intervals to select the interval the observer believes her perceptual decision is more self-consistent than the other. To take into account the perceptual decision in the confidence judgment, we rely on the signed confidence judgment w' described above (Equation 19). We define the choice of the confidence interval C between intervals 1 and 2 as follows

$$C = \arg\max_{i \in \{1,2\}} (\, w_i' \,) \ . \quad (23)$$

According to this definition, the confidence choice will be the interval for which the confidence evidence is the largest in magnitude, except if there is a mismatch between D and D', in which case the confidence choice will be the other interval. The impact of the inconsistency between D and D' is illustrated in Figure 9A. This figure is reproduced from the previous example where the perceptual decisions were R in both intervals (Figure 8). Following Equation 23, interval 1 will be chosen if the confidence evidence lies in the contiguous half space in the lower-right. Applying the confidence decision rule to the other three scenarios of the perceptual decisions in intervals 1 and 2 also leads to contiguous half-spaces that are consistent with a confidence choice in favor of one interval (Figure 9B).

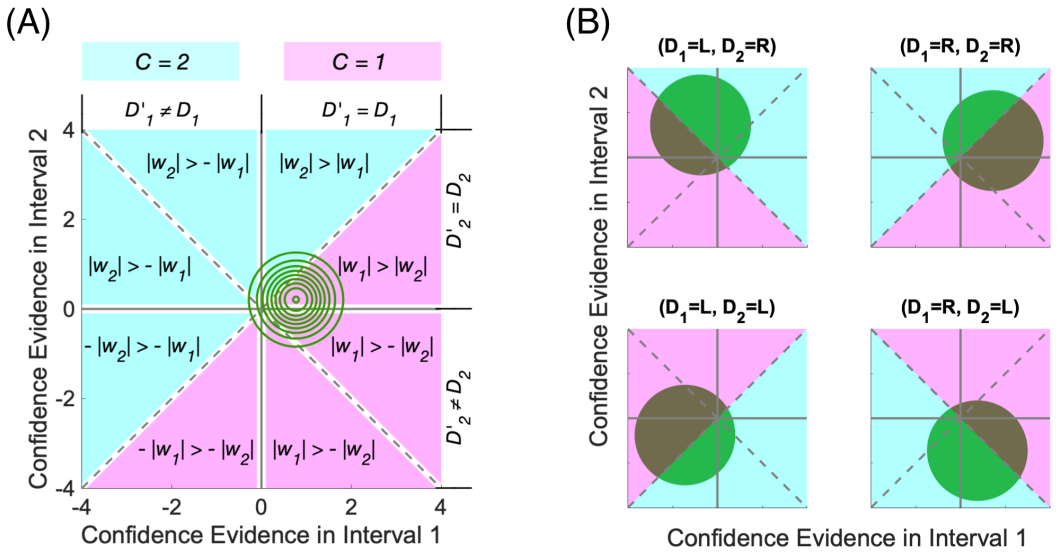

Figure 9. Confidence decision rule. (A) Joint distribution for the confidence evidence conditional on sensory evidence $G(w_1, w_2 \mid s_1, s_2)$ when s_1 and s_2 are both consistent with percept R. This plot is a replica of Figure 8B where eight different sectors are identified from the comparison of the signed confidence evidence across the two intervals. Sectors that lead to choosing interval 1 as more confident are shown in purple ($C = 1$), and those favoring interval 2 in cyan ($C = 2$). Confident choices in favor of interval 1 lie in a contiguous half-space located in the lower-right of the confidence evidence space. (B) Confidence choices for each of the four possible combinations of perceptual decisions across the two intervals. Labels of each panel correspond to the perceptual decisions in each interval (e.g. "$(D_1 = L, D_2 = R)$" indicates that response category L was chosen in interval 1 and R in interval 2). The scenario "$(D_1 = R, D_2 = R)$" illustrated in part (A) of the figure is shown in the upper-right panel.

7.c. Interval Bias

We have to consider one last aspect of the confidence forced-choice paradigm. It is plausible that participants will display some consistent bias in choosing the first or the second interval in all the confidence trials. This type of interval bias has been found to be significant in some individuals, and when it was present, it was relatively stable within individuals (de Gardelle & Mamassian, 2015). If we denote by γ the bias in favor of interval 1, then we can rewrite Equation 23 as follows

$$\begin{cases} C = 1 & \text{if } \; w_1' - w_2' + \gamma > 0 \\ C = 2 & \text{otherwise} \end{cases} . \qquad (24)$$

When there is a bias to choose interval 1 over interval 2 ($\gamma > 0$), interval 1 might be preferred over interval 2 even when the perceptual decision in interval 2 was better than the one in interval 1. This leads to worse discriminability of chosen decisions in interval 1 as compared to interval 2 (Figure 10).

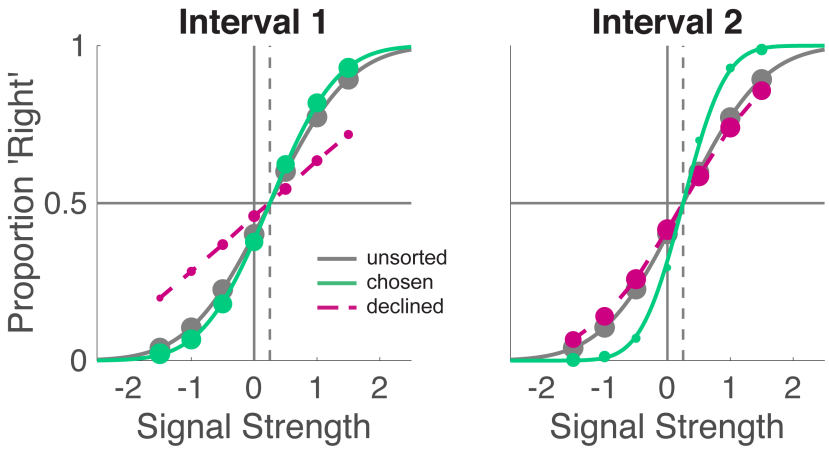

Figure 10. Effect of interval bias on psychometric function. In these simulations, there was a bias in favor of the first interval ($\gamma = 1.0$). The other parameters are listed in Table 2. Plotting conventions are the same as those used in Figure 2B.

The new division of confidence evidence space where intervals 1 and 2 are chosen should take into account this interval bias (Appendix B).

8. Integrated Model for a Confidence Pair

So far, we have considered what is happening on a single confidence pair. In order to make predictions from our model, we need to integrate all possible samples with their respective distributions. This is equivalent to simulating our model with an infinite number of trials.

We start with the joint distribution $G(w_1, w_2 \mid s_1, s_2)$ of confidence evidence conditional on the sensory evidence across the two intervals. Equation 22 provides the mean and covariance of this joint distribution. Following the confidence decision rule, the probability of choosing interval 1 as more confident can be evaluated by integrating over the relevant part of the confidence space, which depends on the perceptual decisions (D_1, D_2) (see Figure 9 and Appendix B). We need to consider separately the four cases corresponding to the 2 by 2 possible perceptual decisions (D_1, D_2). As detailed above, when there is no interval bias ($\gamma = 0$), this space is simply a half-space above or below one of the two diagonals (see Figure 9B). For instance, if both perceptual decisions are 'R' (top-right panel in Figure 9B), we have

$$\mathrm{P}(C = 1 \mid s_1,\ s_2,\ D_1 = \mathrm{R},\ D_2 = \mathrm{R}) = \int_{-\infty}^{+\infty} \int_{-\infty}^{x} G(x, y \mid s_1, s_2)\, dy\, dx \quad . \tag{25}$$

Obviously, the probability of choosing interval 2 as more confident is 1 minus this probability. With a change of variables that rotates the confidence space by $\pi/4$ counter-clockwise, the double integral in Equation 25 can be reduced to a single integral

$$\begin{aligned} &\mathrm{P}(C = 1 \mid s_1,\ s_2,\ D_1 = \text{'R'},\ D_2 = \text{'R'}) \\ &\quad = \int_{-\infty}^{0} \varphi\big(v;\ \big(-Q(s_1;\ \mu_1,\ \sigma_1) + Q(s_2;\ \mu_2,\ \sigma_2)\big)/\sqrt{2}\,, \sigma_c^2\big)\, dv \\ &\quad = \Phi\big(\ \big(Q(s_1;\ \mu_1,\ \sigma_1) - Q(s_2;\ \mu_2,\ \sigma_2)\big)/\big(\sqrt{2}\,\sigma_{\mathrm{c}}\big)\big) \end{aligned} \quad , \tag{26}$$

where $\varphi(x;\ \mu,\ \sigma^2)$ is the probability distribution function of the normal distribution with mean μ and variance σ^2, and Φ is again the cumulative distribution function of the standard normal distribution. We can proceed similarly, for the three other cases to cover all possible pairs of perceptual decisions in intervals 1 and 2,

$$\mathrm{P}(C = 1 \mid s_1,\ s_2,\ D_1,\ D_2) = \begin{cases} \Phi\big(\ \big(Q(s_1;\ \mu_1,\ \sigma_1) - Q(s_2;\ \mu_2,\ \sigma_2)\big)/\big(\sqrt{2}\,\sigma_c\big)\big) & \text{if } D_1 = \text{'R'} \ \&\ D_2 = \text{'R'} \\ \Phi\big(\ \big(Q(s_1;\ \mu_1,\ \sigma_1) + Q(s_2;\ \mu_2,\ \sigma_2)\big)/\big(\sqrt{2}\,\sigma_c\big)\big) & \text{if } D_1 = \text{'R'} \ \&\ D_2 = \text{'L'} \\ \Phi\big(\ \big(-Q(s_1;\ \mu_1,\ \sigma_1) - Q(s_2;\ \mu_2,\ \sigma_2)\big)/\big(\sqrt{2}\,\sigma_c\big)\big) & \text{if } D_1 = \text{'L'} \ \&\ D_2 = \text{'R'} \\ \Phi\big(\ \big(-Q(s_1;\ \mu_1,\ \sigma_1) + Q(s_2;\ \mu_2,\ \sigma_2)\big)/\big(\sqrt{2}\,\sigma_c\big)\big) & \text{if } D_1 = \text{'L'} \ \&\ D_2 = \text{'L'} \end{cases} \quad . \tag{27}$$

When there is an interval bias ($\gamma \neq 0$), these conditional probabilities are still cumulative normal functions, but over a larger or smaller domain (see Appendix B).

When we consider all the possible pairs of sensory evidence presented in the two intervals, we see that the sensory criteria divide the sensory space into four quadrants (see again Equation 2). Applying Equations 27 to the relevant quadrants produces the *confidence choice map* shown in Figure 11B.

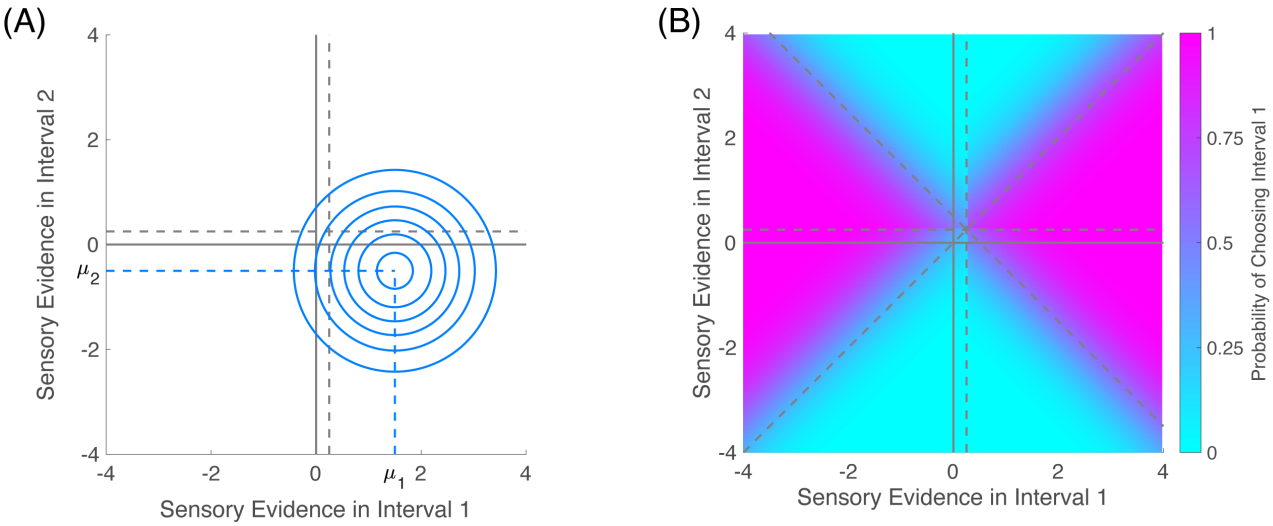

Figure 11. Joint distribution of sensory evidence and confidence choice map. (A) Joint distribution of sensory evidence (replica of Figure 8A reproduced here for convenience). The dashed grey lines indicate the location of the sensory criterion. The dashed blue lines indicate the stimulus strengths in intervals 1 and 2. (B) Confidence choice map. The probability of choosing interval 1 as more confident is plotted for each pair of sensory evidence values in intervals 1 and 2. Parameters for this example are listed in Table 2.

The final step to compute the integrated model is to combine the probability of getting a particular pair of sensory evidence values $(s_1,\ s_2)$ with its associated probability of choosing interval 1 as more confident. The former is the joint distribution of sensory evidences across the two intervals (Figure 11A) and the latter is the confidence choice map (Figure 11B). In layman's terms, we need to multiply point by point Figure 11A with Figure 11B, and then integrate over the whole space.

In formal terms, the probability of choosing interval 1 as more confident is

$$\mathrm{P}(C = 1 \mid D_1,\ D_2) = \frac{\iint_\Omega \mathrm{P}(C = 1 \mid s_1,\ s_2,\ D_1,\ D_2) \cdot \mathrm{P}(s_1, s_2)\, ds_1\, ds_2}{\iint_\Omega \mathrm{P}(s_1, s_2)\ ds_1\, ds_2} \quad , \qquad (28)$$

where Ω is the quadrant of the space of sensory evidence across the two intervals that is compatible with the pair of perceptual decisions $(D_1,\ D_2)$. For instance, when $(D_1,\ D_2) = (\mathrm{R}, \mathrm{R})$, $\Omega = [\theta_s, +\infty) \times [\theta_s, +\infty)$. We can easily compute a numerical approximation for this equation. The result for the different perceptual decisions forms a quadruplet of probabilities as shown in Figure 12A.

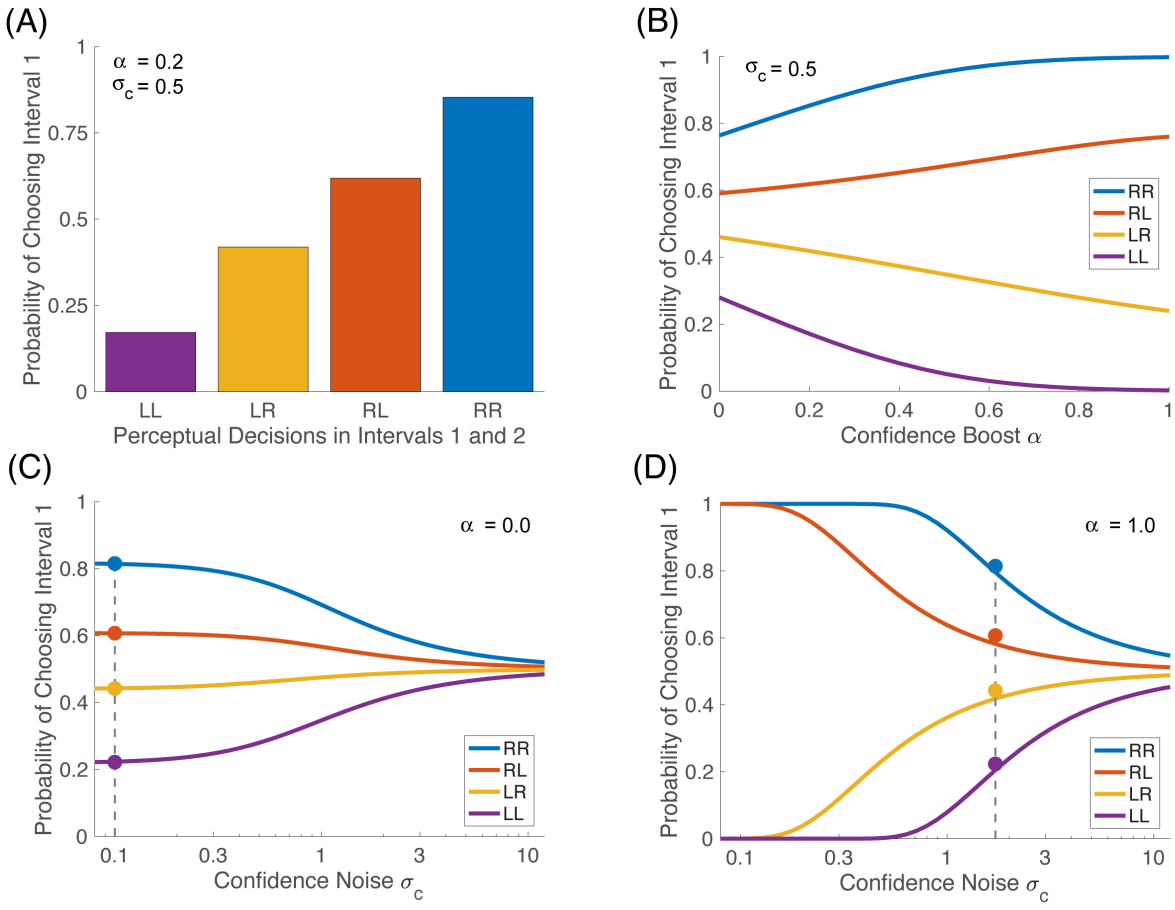

Figure 12. Interval choice probabilities. (A) Quadruplet of confidence choice probabilities for a particular pair of stimuli in the two intervals. The probability of choosing interval 1 as more confidence is plotted for each pair of perceptual decisions in intervals 1 and 2. Labels for the bars correspond to the perceptual decisions in each interval (e.g. "LR" indicates that response category L was chosen in interval 1 and R in interval 2). (B) Effect of confidence boost on interval choice probability. (C) Effect of confidence noise on interval choice probability when the confidence boost is $\alpha = 0$. (D) Effect of confidence noise on interval choice probability when the confidence boost is $\alpha = 1$. The four colored dots in panels (C) and (D) have the same set of four values of interval choice probabilities, therefore the corresponding pairs of confidence boost and confidence noise are equivalent confidence parameters. All parameters, other than the confidence boost in panels (B), (C) and (D), and the confidence noise in panels (C) and (D), are listed in Table 2.

9. Confidence Parameters Indeterminacy and Confidence Efficiency

9.a. Confidence Parameters Indeterminacy

It is instructive to look at the effects of the two main parameters of the model, namely the confidence boost and the confidence noise while keeping the other parameters of the model constant (see Appendix C). Figure 12B illustrates how increasing confidence boost makes the probability of choosing interval 1 deviate from chance level (0.5), for each pair of perceptual decisions. Whether each of these probabilities tends towards 0 or 1 depends on the sign of $|\mu_1 - \theta_s| - |\mu_2 - \theta_s|$ (Appendix C).

Figures 12C and 12D illustrate the effect of confidence noise. As expected, increasing confidence noise makes confidence choices converge towards chance level. This convergence to chance level can be observed both when the confidence boost is small (Figure 12C) and large (Figure 12D).

Comparing Figures 12C and 12D, we can see that confidence boost and confidence noise have opposite effects on interval choice probability. In other words, different pairs of confidence boost and confidence noise trade off and can produce similar outcomes in terms of confidence choice probabilities. One such example is shown with dashed lines in Figures 12C and 12D. These lines indicate that for an arbitrary choice of confidence boost and confidence noise $(\alpha = 0,\ \sigma_c = 0.1)$, one can find other pairs of confidence boost and confidence noise (for instance, $\alpha = 1,\ \sigma_c = 1.71$) that give rise to similar quadruplets of choice probabilities. In other words, estimating both confidence boost and confidence noise is an underdetermined problem that we call *confidence parameters indeterminacy*, and we call *equivalent confidence parameters* the pairs of confidence boost and confidence noise that lead to the same confidence judgments.

We note that confidence parameters indeterminacy is a generic problem for any confidence measurement method, not just the confidence forced-choice paradigm discussed here. In particular, a confidence rating task with a single stimulus strength and a single confidence rating criterion (high vs. low confidence judgments) leads to perfect ambiguity between confidence boost and confidence noise parameters (see Appendix G). One way to beat the curse of confidence parameters indeterminacy is to use multiple stimulus strengths. Increasing the range of stimulus difficulties reduces the uncertainty in the estimation of the confidence boost and confidence noise parameters, and transforms an underdetermined problem into an overdetermined one (see Figure D6 in Appendix D).

Confidence parameters indeterminacy highlights the difficulty in separating out the contribution of confidence boost and confidence noise in confidence judgments. However, in the next section we will see that one can define another metric of confidence performance that combines the contributions of confidence boost and confidence noise.

9.b. Confidence Efficiency

Given quadruplets of confidence choices, sets of equivalent confidence parameters are obtained by choosing the value of confidence boost and searching for the confidence noise that best approximates the confidence choices. Three examples of different sets of equivalent confidence parameters are shown in Figure 13A depicting the trade-off between confidence boost and confidence noise. The set of equivalent confidence parameters corresponding to the ideal confidence observer (blue curve in Figure 13A) is particularly important because it divides the (confidence noise, confidence boost) space into two parts. On its right are all the sets of equivalent confidence parameters that are worse than the ideal confidence observer (green shaded region in Figure 13A), and on its left, the ones that are better (red shaded region). We will come back to this distinction shortly, after defining confidence efficiency.

Sets of equivalent confidence parameters that are better than the ideal confidence observer (e.g. the red curve in Figure 13A) are special because, for these, there exists no confidence noise that can lead to an equivalent confidence performance when the confidence boost is zero. In contrast, note that all sets of equivalent confidence parameters do cross the top horizontal line corresponding to the maximal confidence boost ($\alpha = 1$; horizontal dashed line in Figure 13A). This property allows us to define the *equivalent confidence noise* τ which is the confidence noise of the equivalent confidence parameters that corresponds to $(\alpha = 1)$. These equivalent confidence noises are shown as dots at the top of Figure 13A. The blue dot is the equivalent confidence noise τ_{ideal} for the ideal observer.

The equivalent confidence noise can help us summarize the sensitivity of the confidence judgments for a given set of equivalent confidence parameters, for instance the one shown in green in Figure 13A. We call this summary the *confidence efficiency* η that we define from the inverse of the equivalent confidence noise variance

$$\eta = \tau_{\text{ideal}}^2 \,/\, \tau_{\text{human}}^2 \quad . \tag{29}$$

In this definition, we have normalized the equivalent confidence noise of the human observer by that of the ideal confidence observer, so that the confidence efficiency is exactly 1 for the ideal confidence observer. The ratio of equivalent confidence noises is squared to make confidence

efficiency analogous with the definition of efficiency for perceptual decisions (e.g. Kersten & Mamassian, 2009). Coming back to the two regions of Figure 13A defined by the ideal confidence observer, all sets of equivalent confidence parameters to the right of the curve traced by the ideal confidence observer have a confidence efficiency smaller than 1, and those to its left have a confidence efficiency greater than 1.

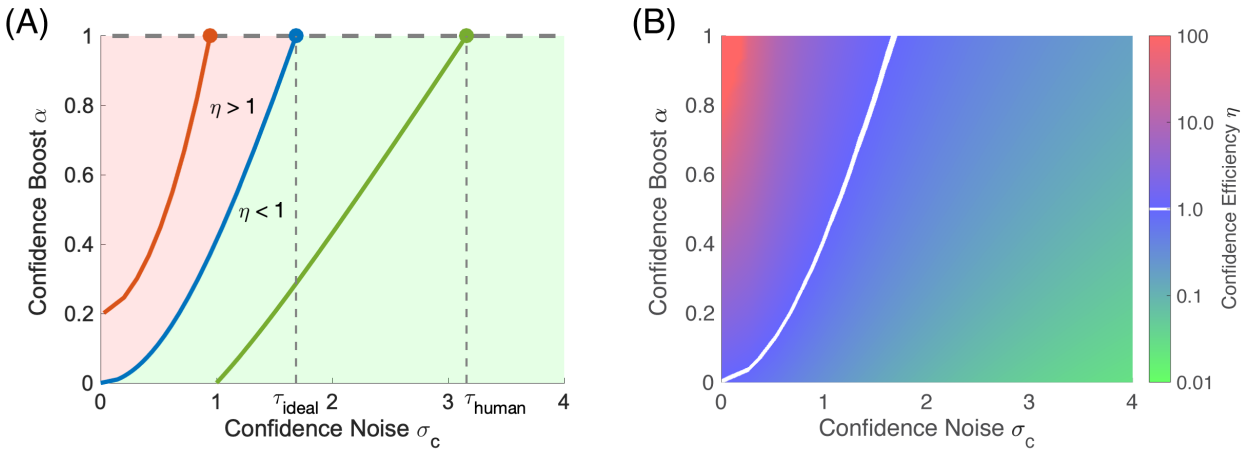

Figure 13. Sets of equivalent confidence parameters and confidence efficiency. (A) Construction of the equivalent confidence noise. Each of the three colored curves shows one set of equivalent confidence parameters, namely the pairs of confidence noise and confidence boost that produce similar quadruplets of choice probabilities across all four possible perceptual decisions of a confidence pair. The blue curve corresponds to the ideal confidence observer ($\alpha = 0,\ \sigma_c = 0$). It intersects the line of maximal confidence boost ($\alpha = 1$; horizontal dashed line at the top) at a point called the equivalent confidence noise for the ideal confidence observer (τ_{ideal}). For each set of equivalent confidence parameters, we can similarly find the equivalent confidence noise (e.g. the value τ_{human} for the green curve that corresponds to a noisy ideal confidence observer ($\alpha = 0,\ \sigma_c = 1$)). (B) Confidence efficiency. The equivalent confidence noise can be used to compute the confidence efficiency (for the green set of equivalent confidence parameters in panel (A), the efficiency is $\eta = 0.285$). By definition, confidence efficiency is 1 when both confidence boost and confidence noise are null. Confidence efficiency increases with confidence boost and decreases with confidence noise. Any pair of confidence noise and confidence boost that are to the right and below of the blue curve in panel (A) have a confidence efficiency smaller than 1, and those to the left and above have a confidence efficiency greater than 1.

Using our definition of confidence efficiency, we can assign a confidence efficiency for each pair of confidence noise and confidence boost (Figure 13B). Confidence efficiency runs from zero (no metacognition, obtained when confidence noise is very large) to infinity (super-ideal confidence observer, obtained when confidence boost is 1 and there is no confidence noise). By definition, confidence efficiency is 1 for all the pairs of confidence noise and confidence boost that are equivalent confidence parameters of the ideal confidence observer.

10. Full Model and Parameters Estimation

When we introduced the problem of confidence parameters indeterminacy in the previous section, we highlighted that confidence boost and confidence noise were difficult to estimate simultaneously. There are however small differences in the quadruplets of choice probabilities for different pairs of these parameters (compare again Figure 12A with Figure 12B) as long as the dataset contains comparisons across different stimulus strengths (see Figure D6 in Appendix D). In that case, there will be a pair of confidence boost and confidence noise that best explains all the choice probabilities.

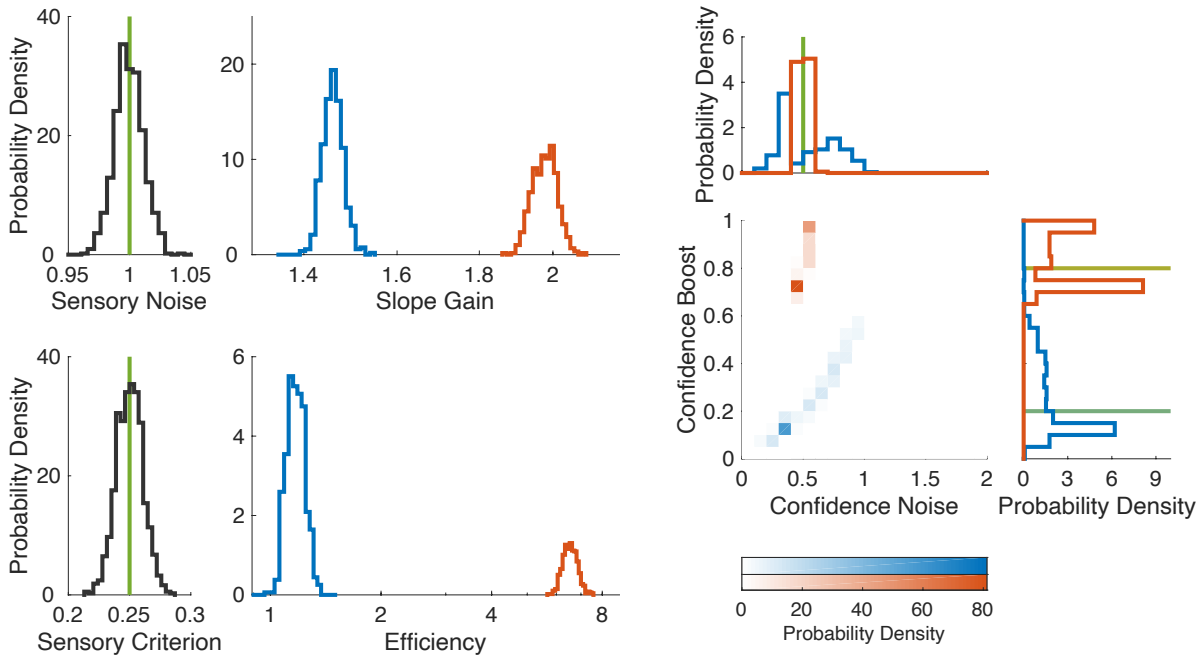

Figure 14. Parameter recovery of the model. The distributions of parameters were estimated from 500 simulated experiments. The estimated parameters were the sensory noise σ_s and the sensory criterion θ_s (first column), the gain in the slope of the psychometric functions between chosen and unsorted trials and the confidence

efficiency (second column). The full confidence model also attempted to infer the confidence boost α and the confidence noise σ_c (right panel). Estimated confidence boost and confidence noise are correlated, and this correlation is at the origin of the problem of confidence parameters indeterminacy. The original parameter values that were used in the simulations are shown as green lines. Two different confidence boosts were simulated, $\alpha = 0.2$ in blue and $\alpha = 0.8$ in orange. The other parameters are listed in Table 2.

Assuming that the confidence pairs are independent from each other, we can obtain the set of best model parameters by summing the log likelihood of each confidence pair. An example of best fitted estimate is shown superimposed on the simulated data in Figure 3. In that figure, simulated parameters were $\sigma_s = 1.0$, $\sigma_c = 0.5$, $\theta_s = 0.25$, and $\alpha = 0.2$. Estimated parameters were $\hat{\sigma}_s = 0.999$, $\hat{\sigma}_c = 0.326$, $\hat{\theta}_s = 0.245$, $\hat{\alpha} = 0.122$, and $\hat{\eta} = 0.789$ (θ_c, β, and γ were fixed to their default values). We see that estimated parameters are near their theoretical values, but there are small deviations.

To appreciate the faithfulness of our model parameters, we simulated 500 experiments with the same original parameters, and collected the distributions of the estimated parameters. Figure 14 shows these distributions for two different values of confidence boost ($\alpha = 0.2$ vs. $\alpha = 0.8$). We observe that these two values of confidence boost can be distinguished since their distributions do not overlap. In addition, both the gain in the slope of the psychometric functions and the efficiency measures are able to distinguish these two conditions, since the distributions are clearly segregated (middle column of Figure 14).

The next figure shows simulations of the model with varying levels of confidence noise or varying levels of confidence boost (Figure 15). Critically, the estimated confidence noise follows very well the actual confidence noise for the two levels of confidence boost simulated (Figure 15A, top), and these levels of confidence boost are well-recovered independently of the confidence noise (Figure 15A middle). The opposite holds when varying the confidence boost: the estimated boost tracks the simulated boost, both for simulations with high confidence noise and low confidence noise (Figure 15B middle), with a possible exception when confidence efficiency is extremely high (combination of high confidence boost and low confidence noise). In addition, the two simulated levels of confidence boost are well-recovered for any value of the confidence noise (Figure 15A middle) and the two simulated levels of confidence noise are well-recovered for any value of the confidence boost (Figure 15B top). In short, both confidence noise and confidence boosts can be recovered very well, at least within a reasonable range of values.

In Appendix D, we present parameter recovery for the remaining parameters of the model. The confidence noise and boost parameters are quite stable for different values of sensory noise. This is not surprising since, in the model, confidence evidence is normalized by sensory sensitivity, so the confidence noise and boost parameters should not depend on sensory noise. The confidence noise and boost parameters are also quite stable for different values of sensory and confidence criteria, at least as long as these criteria are within reasonable limits of the range of the presented sensory stimuli. Importantly, the confidence noise and boost parameters are very stable for different values of biases in favor of responding either the first or second interval. In this latter case though, confidence efficiency decreases as the interval response bias increases, because favoring one interval over the other necessarily impairs the accuracy of choosing the interval that was more likely to be self-consistent. Finally, the confidence noise and boost parameters are better recovered as more confidence pairs are tested in an experiment. If the number of confidence pairs is less than about 1,000, the confidence noise and boost parameters are estimated too imprecisely, but importantly the confidence efficiency remains a robust measure of meta-perception.

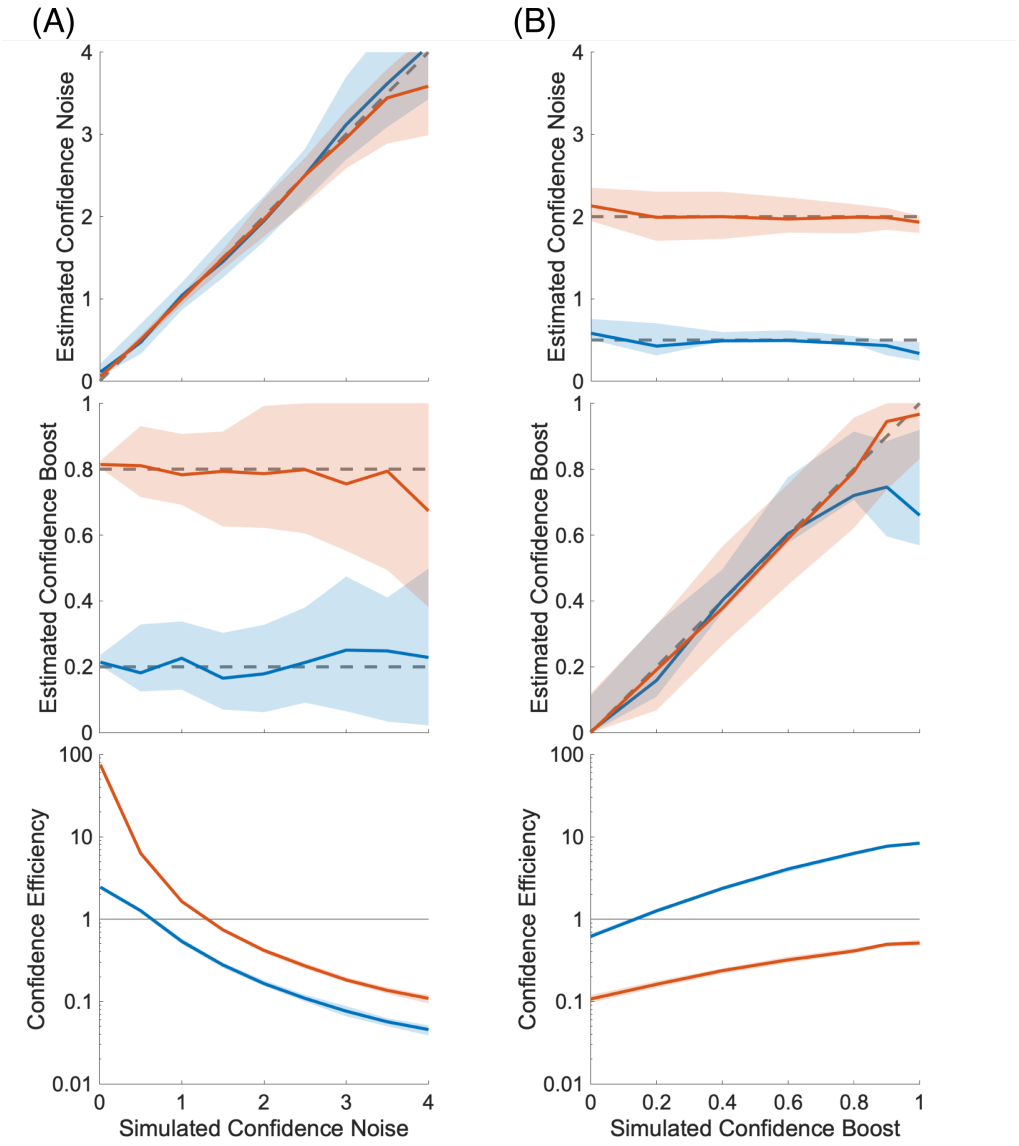

Figure 15. Model recovery for a range of confidence noise and confidence boost. (A) The plots show estimated parameters for two different values of confidence boost, $\alpha = 0.2$ in blue and $\alpha = 0.8$ in red. The estimated parameters are confidence noise (top panel), confidence boost (middle), and efficiency (bottom). (B) The plots show estimated parameters for two different values of confidence noise, $\sigma_c = 0.5$ in blue and $\sigma_c = 2.0$ in red. The thick lines are median estimated values across $N = 100$ repeated simulations, and the shaded areas cover the 25th to the 75th interquartile range.

At this stage, we have not presented the model recovery for the last parameter of the model, the confidence bias β. This is because this scaling factor affects both intervals equally, so its effects cancel out in the confidence forced-choice paradigm (see section 5.c). In a sense, the confidence forced-choice paradigm was designed to be immune to possible confidence biases, so it was

expected that this bias would be difficult to estimate. However, there is one scenario where the confidence bias can be recovered, at least up to a scaling factor, and this is what we explore next.

11. Effects of Confidence Bias

So far, we have considered that participants were performing the same perceptual task in both intervals of a confidence pair. However, it is interesting to consider the condition where the participant is asked to perform different tasks across the two intervals. This condition allowed us to claim that confidence was computed in a common currency, rather than in some metric that is tightly constrained by the dimension along which the task is performed (de Gardelle & Mamassian, 2014; de Gardelle, Le Corre, & Mamassian, 2016).

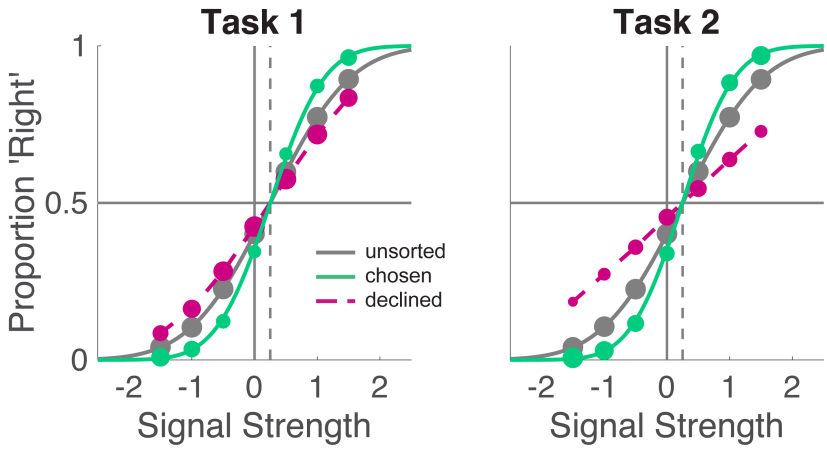

Figure 16. Effect of confidence bias on the psychometric functions. In these simulations, the first task was properly scaled ($\beta = 1.0$) but the observer was over-confident in the second task ($\beta = 2.0$). As a result, whenever task 1 is competing with task 2 in a confidence pair, confidence choice is biased in favor of task 2 (indicated by larger green dots for task 2 than for task 1). All parameters except β are identical across the two tasks and listed in Table 2. Plotting conventions are the same as those used in Figure 2B.

A between-task confidence judgment also allows us to tackle an issue that we had to leave out when participants were performing the same task in both intervals of a confidence pair. This issue is whether participants are properly estimating their perceptual sensitivity in a task and correctly using this estimate to normalize their confidence evidence. In the model we described above, we assumed that this normalizing parameter β was indeed 1.0 (no confidence bias). If only one task is

used, the effects of this parameter are invisible (Figure 6C is identical to Figure 2B), because the same scaling is applied to both confidence evidences of the two intervals. When two tasks are competing in the two intervals, the psychometric functions for the two tasks sorted by confidence judgments are now different (Figure 16). In general, though, comparing psychometric functions across tasks makes little sense because the units of stimulus strengths will be different. However, we will be able to run our model and try to best predict the probabilities of choosing with confidence one task over the other across all stimulus strengths and perceptual decisions. When two tasks are run, we cannot estimate both corresponding β parameters, but we can estimate their ratios (see Appendix E). This allows us to estimate whether one task shows over- or under-confidence relative to the other task.

12. Re-Analysis of De Gardelle & Mamassian (2015)

So far, we have looked at the ability of the model to simulate a confidence forced-choice experiment, and the faithfulness of the recovered parameters. We now apply this framework to the re-analysis of one of our previous studies. We choose the study of confidence for motion direction discrimination that was published in de Gardelle & Mamassian (2015). In that experiment, observers had to discriminate the mean direction of motion above or below a reference for a stimulus composed of multiple random dot motion. The strength of the stimulus was manipulated by varying the mean motion direction, where larger mean motion directions away from the reference are easier stimuli. In addition, there were two stimulus uncertainty levels, represented by the different ranges of motion directions of the dots within a stimulus. Given that these ranges are very different, we can apply the analysis of confidence biases that we discussed in section 11, where the two tasks correspond here to the two stimulus uncertainty levels.

We present here parameter estimates based on the group data. This group data set corresponds to the data collected across all participants, after normalizing each participant to her own sensory noise and criterion. The analysis thus assumes that there is single set of model parameters shared across all participants. In this sense, this analysis can be seen as complementary to the one presented in the original paper (de Gardelle & Mamassian, 2015), where individual differences were emphasized.

Parameter estimates for this experiment are shown in Figure 17. Confidence efficiency was about 0.5, indicating that participants were clearly able to make meta-perceptual judgments (efficiency larger than 0) but less efficient than the ideal confidence observer (efficiency less than 1). Separating confidence efficiency into confidence noise and confidence boost, we found evidence that confidence in this task and for this stimulus was relying more on primary than secondary confidence evidence (confidence boost closer to zero than to one). We used likelihood ratio tests

for nested models (Mood, Graybill, & Boes, 1974) to test whether confidence boost was either zero or one. We computed the test statistic $\lambda_{LR} = -2(\lambda_0 - \lambda_1)$, where λ_0 is the log-likelihood of the constrained model where confidence boost is fixed to 0 (or 1), and λ_1 is the log-likelihood of the unconstrained model where confidence boost is free to vary. If the model where confidence boost is fixed is correct, then this test statistic is asymptotically distributed as a χ_1^2 random variable (the degrees of freedom is 1 because the two models differ by only one parameter). For the pooled participant, we could not reject the hypothesis that confidence boost was 0 ($\chi_1^2 = 1.69, \mathrm{p} = 0.194$), but we could reject the hypothesis that confidence boost was 1 ($\chi_1^2 = 13.4, \mathrm{p} < 0.001$).

Confidence noise was estimated to be about 1 (this value does not have any unit and thus could potentially be compared to other confidence noise in other experiments). Finally, we also found a small but significant confidence bias, revealing an overconfidence for the high stimulus uncertainty relative to low stimulus uncertainty. In other words, on average, participants did not fully appreciate the effect of the stimulus noise on their sensory sensitivity.

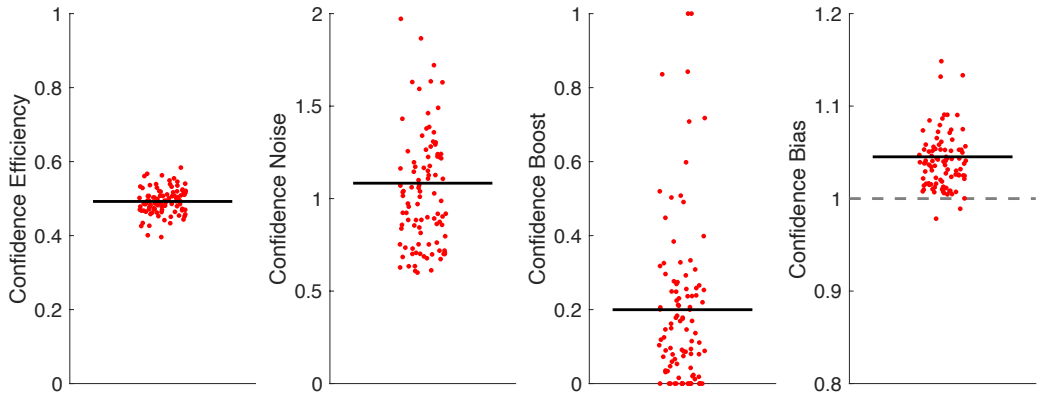

Figure 17. Model parameter estimates in a real study. Individual dots are estimates from 100 bootstrapped trials on the data collected over 15 observers. Data are from de Gardelle & Mamassian (2015).

Our reported confidence efficiency is smaller than what is typically reported in the literature. At this stage, it would be hazardous to compare our confidence efficiency measure to the meta-d'/d' approach (Maniscalco & Lau, 2012) given that we reported confidence efficiency for a single experiment. And before making this comparison, we should first note that our efficiency is based on a ratio of variances, whereas meta-d'/d' involves a ratio of standard deviations. When this is taken into account, the meta-d'/d' value of 0.77 reported initially by Maniscalco & Lau (2012) would amount to a ratio of variance of 0.59, which is not very different from our estimate of 0.5 reported here. Residual differences between our confidence efficiency measure and the meta-d'/d'

approach can have different origins. These include different experimental procedures for the perceptual judgment, different experimental procedures for the confidence judgment, and different populations of participants. It will be worth computing this new efficiency measure over a large range of studies to better compare the two efficiency measures.

13. Discussion

In summary, we have presented here a generative model for the estimation of confidence in perceptual decisions. Our model considers confidence to be the evaluation that one's perceptual decision is self-consistent, thereby highlighting that confidence is about a decision, not about the stimulus itself, its sensory uncertainty, contrast, duration or visibility. The self-consistency aspect of the definition emphasizes that the perceiver evaluates her own percept, rather than whether her percept is consistent with the true state of the world. Using this definition, we have proposed a model of perceptual confidence where the perceptual decision follows classical Signal Detection Theory (Green & Swets, 1966). We then assumed that confidence evidence scales with the distance between sensory evidence and the sensory criterion, where the scaling factor is inversely proportional to sensory noise. This confidence evidence is corrupted by confidence noise but can benefit from some confidence boost that corresponds to the possibility that confidence may rely on additional information compared to the sensory evidence. We identify three keys aspects by which our approach goes beyond previous work.

First, we can theoretically differentiate between primary and secondary sources of information for the computation of confidence. To obtain this result, we described the behavior of an ideal agent that uses the same information as that used for the perceptual decision. This ideal confidence observer was contrasted to a super-ideal agent that uses a novel and perfect estimation of the stimulus for the purpose of the confidence judgment. An observer that relies exclusively on primary confidence evidence mimics the ideal confidence observer, albeit not optimally (see also Bang et al., 2019), whereas an observer that relies only on secondary evidence mimics the super-ideal confidence observer. The fraction of ideal and super-ideal observers in the confidence judgments is represented by the confidence boost parameter in our model. To be precise, this parameter reflects the fraction of all information used in confidence processing that was not used for the perceptual decision (see also Barrett, Dienes & Seth, 2013; Maniscalco & Lau, 2016; Fleming & Daw, 2017). As such, it may aggregate information from multiple sources, including non-sensory information such as motor signals (see e.g. Fleming et al., 2015; Wokke et al., 2020), or fluctuations of attention (see e.g. Recht et al., 2019), or sensory information that was processed after the perceptual decision took place (Baranski & Petrusic, 1998; Pleskac & Busemeyer, 2010). Similarly, we should emphasize that the noise corrupting the confidence evidence, although

quantified with a single parameter in our model, may aggregate multiple sources of inefficiencies, including noisy read-out of the perceptual evidence, but also influences from previous confidence judgments (Rahnev et al., 2015), or influences from other features that are not related to perceptual performance. Importantly, the confidence boost parameter was well recovered in our simulations that contained a large number of trials. We anticipate that the ability to distinguish between primary and secondary sources of confidence evidence will be an important asset of our model.

Second, we propose a measure of efficiency that is genuinely anchored to the metacognitive level of computation. Our efficiency measure is obtained by comparing human confidence performance to that of the ideal confidence observer. The introduction of our confidence efficiency is motivated by a problem of confidence parameters indeterminacy between two parameters of our model, confidence noise and confidence boost. Both of these parameters are hard to differentiate in an experiment that contains only a limited range of stimulus strengths or a limited number of trials, but can be estimated when confidence is measured for multiple stimulus strengths. Our metric of confidence efficiency captures the trade-off between confidence boost and confidence noise, and importantly, it can be reliably estimated even when confidence is measured for a single stimulus strength. Our definition of confidence efficiency deviates from previous ones. For instance, in the now popular *meta-d'* framework for analyzing confidence judgments (Maniscalco & Lau, 2012), no generative model is specified for confidence judgments. Under that framework, *meta-d'* quantifies the sensitivity at the metacognitive level by estimating the first-order sensitivity that would be needed to observe the data if the metacognitive system were perfect. The *M-ratio*, that is the ratio of *meta-d'* over *d'*, has been put forward as a measure of efficiency, but although it makes some intuitive sense, it does not correspond to a clear process. Other theoretical approaches to metacognition have described potential generative models for confidence judgments (e.g. Pleskac & Busemeyer, 2010; Sanders et al., 2016; Fleming & Daw, 2017), but they did not offer an efficiency measure based on these models.

Third, our model can sometimes recover the confidence bias that corresponds to the mis-estimation of one's perceptual sensitivity. In our model, perceptual sensitivity is used to normalize confidence so that this latter can be compared across tasks and sensory modalities (de Gardelle & Mamassian, 2014; de Gardelle et al., 2016). As a consequence, overconfidence corresponds here to an over-estimation of one's perceptual sensitivity. While the effects of confidence bias are invisible when one considers only one task, the ratio of confidence biases can be estimated when two tasks are compared.

Confidence comparison between two tasks is particularly easy within the confidence forced-choice paradigm. In this paradigm, a confidence choice is taken between two perceptual decisions. Using our modelling framework, we have described the probabilities with which one perceptual decision

is associated with a larger confidence than the other decision, for different stimulus strengths and different commitments to perceptual decisions. Previous analyses of metacognitive abilities have had troubles to take into account varying difficulty levels. For instance, the classic measure of confidence resolution simply compares confidence in correct responses and errors, and ignores task difficulty. In the *meta-d'* approach, one major limitation is that it is designed to analyze data where perceptual sensitivity is constant across trials (only one stimulus strength is used in the experiment). Failure to meet this assumption leads to overestimations of metacognitive sensitivity (see e.g. Rahnev & Fleming, 2019), because participants could be using variations of performance that cannot be used in the *meta-d'* estimation procedure. Our confidence forced-choice method may allow researchers to overcome this obstacle.

Our model involves a number of parameters and assumptions, which deserve scrutiny. We argue however that most assumptions of our model are relatively standard and supported in part by empirical evidence. Besides, the parameters we have introduced all have a clear interpretation, and can be recovered quite well (see section 10 and Appendices D and E). The output of our model is a signed confidence evidence that approximates the probability that the perceptual decision is self-consistent. When applied to the confidence forced-choice paradigm, the decision rule for confidence is a simple comparison of the signed confidence evidence between two trials, and does not involve any complex inference. In this respect, our approach appears less demanding than the *actor-critic* model of Fleming & Daw (2017). The reason is that, in our model, confidence evidence is obtained from sensory evidence in a simple processing step that involves some scaling and some additive noise. In contrast, in the model of Fleming & Daw (2017), confidence evidence is used to infer the full distribution of sensory evidence that would be compatible with this confidence evidence. This inferred sensory evidence does not necessarily match the actual sensory evidence, and this could explain potentially interesting paradoxical effects of action on confidence (e.g. Pereira et al., 2020). However, this inference has a cost, that of knowing the covariance between confidence evidence and sensory evidence. It is arguably unrealistic to assume that human participants have access to this latter knowledge, and it becomes computationally intense when multiple levels of difficulty are involved.

We believe that our model could also be pertinent for the interpretation of data obtained with a more traditional confidence rating paradigm. In Appendix G, we show the initial steps to use the model with confidence rating data. The Appendix shows that for a single stimulus strength, we are faced with the problem of confidence parameters indeterminacy so that the confidence boost and confidence noise parameters cannot be jointly estimated. Presumably, these parameters can be estimated when multiple stimulus strengths are presented to the observer, but where to place optimally the confidence criteria becomes a serious issue. Further work is clearly necessary here.

One aspect of our model that appears non-trivial is the possibility that participants would use distinct decision criteria for the Type 1 response and for the Type 2 evaluation. This possibility was explicitly excluded in the *meta-d'* framework. Our framework allows for it, although we anticipate that a reduced model without this additional criterion should suffice in most case. However, this parameter might be interesting to researchers in some situations, where participants have to combine sensory and non-sensory information about a stimulus. The non-sensory information can be a probabilistic cue, as in many decision-making studies (e.g. Locke et al., 2020), or an advice given by another observer, as for instance in Asch's conformity experiment (Asch 1956). Here, as they face a tradeoff between optimality and accuracy, participants might use a Type 1 criterion that takes into account all the cues to make their own decision, but a Type 2 criterion that only considers their own sensory information when evaluating their confidence. Future research, both theoretical and empirical, may aim at understanding how metacognition unfolds in these situations of decision under influence.

To conclude, our effort has focused on specifying a formal generative model where confidence can be both corrupted and boosted relative to the sensory evidence, and the application of this model to the confidence forced choice paradigm. Obviously, this generative model could be used to derive confidence ratings on a scale, which are most commonly used in experiments. Doing so would require introducing additional parameters for the mapping between internal and reported confidence (Aitchison et al., 2015), which the confidence forced choice paradigm naturally avoids. One other direction for future work is to extend the present model to other perceptual tasks, including detection tasks (see e.g. García-Pérez et al., 2011). Finally, since the simultaneous estimation of all parameters in our model require a large amount of data, the development of a Bayesian hierarchical estimation would be important to be able to collapse data across participants (Fleming, 2017). Ultimately, it will be interesting to compare the parameters of the generative model across tasks, sensory modalities, and participant populations.

Acknowledgments

This work was supported by grants ANR-10-BLAN-1910 "Visual Confidence" to PM, and ANR-18-CE28-0015 "VICONTE" to PM and VdG. Supplementary support came from ANR-17-EURE-0017. The authors would like to thank Tarryn Balsdon, Thibault Gajdos, Mike Landy, Shannon Locke, Jérôme Sackur, and four meticulous anonymous reviewers for their critical comments on an earlier draft of the manuscript.

Source code in Matlab for model fitting is available at: https://github.com/mamassian/cfc

References

Adler, W. T., & Ma, W. J. (2018). Limitations of Proposed Signatures of Bayesian Confidence. *Neural Computation*, 1–28.

Aguilar-Lleyda, D., Lemarchand, M., & de Gardelle, V. (2020). Confidence as a priority signal. *Psychological Science*, *31*(9), 1084-1096.

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Computational Biology*, *11*(10), e1004519. http://doi.org/10.1371/journal.pcbi.1004519

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70.

Balsdon, T., Wyart, V., & Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature Communications*, *11*(1), 1753–11. http://doi.org/10.1038/s41467-020-15561-w

Bang, J. W., Shekhar, M., & Rahnev, D. A. (2019). Sensory noise increases metacognitive efficiency. *Journal of Experimental Psychology: General*, *148*(3), 437–452.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 929–945.

Barlow, H. B. (1962). A method of determining the overall quantum efficiency of visual discriminations. *The Journal of Physiology*, *160*(1), 155–168.

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–552.

Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. *PLoS Computational Biology*, *5*(9), 1–8. http://doi.org/10.1371/journal.pcbi.1000504

Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance, 43*(8), 1520-1531.

Bosch, E., Fritsche, M., Ehinger, B. V., & de Lange, F. P. (2020). Opposite effects of choice history and evidence history resolve a paradox of sequential choice bias. *Journal of Vision*, *20*(12), 9–9. http://doi.org/10.1167/jov.20.12.9

Clarke, F. R., Birdsall, T. G., & Tanner, W. P., Jr. (1959). Two types of ROC curves and definitions of parameters. *The Journal of the Acoustical Society of America*, *31*, 629–630.

de Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a common currency between vision and audition. *PLoS ONE*, *11*(1), e0147901. http://doi.org/10.1371/journal.pone.0147901

de Gardelle, V., & Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? *Psychological Science*, *25*(6), 1286–1288.

de Gardelle, V., & Mamassian, P. (2015). Weighting mean and variability during confidence judgments. *PLoS ONE*, *10*(3), e0120870. http://doi.org/10.1371/journal.pone.0120870

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature neuroscience*, *16*(1), 105-110.

Fleming, S. M. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *3*(1:nix007), 1–14. http://doi.org/10.1093/nc/nix007

Fleming, S. M., & 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., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280–1286.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 1–9. http://doi.org/10.3389/fnhum.2014.00443

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. C. (2015). Action-specific disruption of perceptual confidence. *Psychological Science*, *26*(1), 89–98.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–876.

García-Pérez, M. A., Alcalá-Quintana, R., Woods, R. L., & Peli, E. (2011). Psychometric functions for detection and discrimination with and without flankers. *Attention, Perception & Psychophysics*, *73*(3), 829–853.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, *96*(2), 267–314.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Hainguerlot, M., Vergnaud, J. C., & de Gardelle, V. (2018). Metacognitive ability predicts learning cue-stimulus associations in the absence of external feedback. *Scientific Reports*, *8*(1), 1–8.

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review*, *119*(1), 186–200.

Kersten, D., & Mamassian, P. (2009). Ideal observer theory. In L. R. Squire (Ed.), *Encyclopedia of Neuroscience*, volume 5 (pp. 89-95). Oxford: Academic Press.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113.

Locke, S. M., Gaffin-Cahn, E., Hosseinizaveh, N., Mamassian, P., & Landy, M. S. (2020). Priors and payoffs in confidence judgments. *Attention, Perception & Psychophysics*, *77*(2), 638–18.

Mamassian, P. (2016). Visual confidence. *Annual Review of Vision Science*, *2*(1), 459–481.

Mamassian, P. (2020). Confidence forced-choice and other metaperceptual tasks. *Perception*, *49*(6), 616–635.

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

Maniscalco, B., & Lau, H. C. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness*, *2016*(1):niw002, 1–17. http://doi.org/10.1093/nc/niw002

Maniscalco, B., Peters, M. A. K., & Lau, H. C. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception & Psychophysics*, *78*(3), 923–937.

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

Mood, A., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the Theory of Statistics* (3rd edition). New York: McGraw-Hill.

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

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

Moreira, C. M., Rollwage, M., Kaduk, K., Wilke, M., & Kagan, I. (2018). Post-decision wagering after perceptual judgments reveals bi-directional certainty readouts. *Cognition*, *176*, 40–52.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, *26*, 125–173.

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

Pereira, M., Faivre, N., Iturrate, I., Wirthlin, M., Serafini, L., Martin, S., et al. (2020). Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging. *Proceedings of the National Academy of Sciences*, *117*(15), 8382–8390.

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

Rahnev, D. A., & Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness*, *2019*(1), niz009. http://doi.org/10.1093/nc/niz009

Rahnev, D. A., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. C. (2015). Confidence leak in perceptual decision making. *Psychological Science*, *26*(11), 1664–1680.

Rausch, M., & Zehetleitner, M. (2016). Visibility is not equivalent to confidence in a low contrast orientation discrimination task. *Frontiers in Psychology,* *7*:591, 1–15. http://doi.org/10.3389/fpsyg.2016.00591

Rausch, M., & Zehetleitner, M. (2019). The folded X-pattern is not necessarily a statistical signature of decision confidence. *PLoS Computational Biology*, *15*(10), e1007456. http://doi.org/10.1371/journal.pcbi.1007456

Recht, S., Mamassian, P., & de Gardelle, V. (2019). Temporal attention causes systematic biases in visual confidence. *Scientific Reports*, *9*:11622, 1–9. http://doi.org/10.1038/s41598-019-48063-x

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266.

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

Scott, R. B., Dienes, Z., Barrett, A. B., Bor, D., & Seth, A. K. (2014). Blind insight: Metacognitive discrimination despite chance task performance. *Psychological Science*, *25*(12), 2199–2208.

Spence, M. L., Dux, P. E., & Arnold, D. (2016). Computations underlying confidence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(5), 671–682.

van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). Confidence is the bridge between multi-stage decisions. *Current Biology*, *26*(23), 3157-3168.

Wilimzig, C., Tsuchiya, N., Fahle, M., Einhäuser, W., & Koch, C. (2008). Spatial attention increases performance but not subjective confidence in a discrimination task. *Journal of Vision*, *8*(5), 7.1–10. http://doi.org/10.1167/8.5.7

Wokke, M. E., Achoui, D., & Cleeremans, A. (2020). Action information contributes to metacognitive decision-making. *Scientific Reports*, *10*: 3632, 1–15. http://doi.org/10.1038/s41598-020-60382-y

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, *367*(1594), 1310–1321.

Zizlsperger, L., Sauvigny, T., & Haarmeier, T. (2012). Selective attention increases choice certainty in human decision making. *PLoS ONE*, *7*(7): e41136. https://doi.org/10.1371/journal.pone.0041136

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*, 79. http://doi.org/10.3389/fnint.2012.00079