Title
A measure of metacognitive bias you can be confident in

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors
O'Neill, Kevin
Zhan, Tianqi
Fleming, Stephen M

Publication Date
2026

A measure of metacognitive bias you can be confident in

Kevin O'Neill, Tianqi Zhan & Stephen M. Fleming
{kevin.o'neill, tianqi.zhan.23, stephen.fleming}@ucl.ac.uk
Department of Experimental Psychology and Institute of Cognitive Neuroscience, University College London

Abstract

Research on metacognition critically relies on psychophysical measures that are independent of task performance. Although there exist well-validated measures of *metacognitive efficiency* (i.e., the degree to which confidence ratings distinguish between correct and incorrect decisions), relatively less attention has been paid to developing corresponding measures of *metacognitive bias* (i.e., the overall tendency to make responses with high confidence). Most researchers currently use mean confidence as a proxy for metacognitive bias, however this measure is confounded by changes in accuracy, response bias, and metacognitive sensitivity. Here, we introduce a new signal detection theoretic measure of metacognitive bias (meta-Δ) and demonstrate that it is independent of a range of potential confounds. Then, by analyzing longitudinal data, we show that this measure is reliable enough to use in psychological research. Overall, we advocate for the use of meta-Δ as a principled measure of metacognitive bias.

Keywords: metacognition; confidence; signal detection theory

Introduction

People differ greatly in how they report their confidence in their decisions: politicians, for example, have a strong incentive to make decisions with high confidence. One's overall level of (over- or under-) confidence in individual decisions is not only predictive of behaviors such as collective decision-making (Bang et al., 2017, 2020; Hertz et al., 2017) and information-seeking (Schulz et al., 2020; Van Marcke & Desender, 2025), but is also foundational to broader self-beliefs (Katyal et al., 2025; Seow et al., 2021) and consequently associated with differences in mental health (Hoven et al., 2019; Katyal & Fleming, 2026; Rouault et al., 2018).

Modern psychological research assesses metacognition by relating confidence ratings to actual task performance. Just as psychophysicists develop comprehensive models to distinguish performance factors from response characteristics, metacognition researchers strive for a clean distinction between *metacognitive sensitivity*—the degree to which confidence tracks accuracy—and *metacognitive bias*—the overall level of confidence (Fleming & Lau, 2014). In particular, while there remains some debate about the success of this endeavor (e.g., Guggenmos, 2021; Rausch et al., 2023; Xue et al., 2021), a substantial degree of progress has been made developing measures of metacognitive sensitivity that are independent of task performance and metacognitive bias (Fleming & Lau, 2014; Maniscalco & Lau, 2012, 2014; see also Boundy-Singer et al., 2023; Mamassian & de Gardelle, 2025).

In comparison, researchers have focused less on developing independent measures of metacognitive bias. The most prominent measure of metacognitive bias is simply mean confidence (Fleming & Lau, 2014; Rahnev, 2025). Insofar as metacognitive bias is conceptualized as one's overall confidence level, mean confidence is a natural choice: it is clear that differences in metacognitive bias should result in differences in mean confidence. However, mean confidence is also confounded by other aspects of task performance, making it an imprecise guide to the subject's intrinsic tendency towards over- or underconfidence. For instance, Sherman et al. (2018) found that mean confidence is sensitive to task-level features such as accuracy and response bias, as well as metacognitive sensitivity. But in the absence of a clearly better alternative, mean confidence remains popular. As a result, it is pertinent to develop new measures of metacognitive bias that control for the influence of other aspects of task performance.

After a brief review of signal detection theoretic models of metacognitive sensitivity, we introduce a new measure of metacognitive bias, meta-Δ. As meta-Δ is derived from a commonly used model of metacognitive sensitivity (the meta-d' model), this bias measure can readily be incorporated into the standard workflow in metacognition research. We show in simulation that meta-Δ is sensitive and specific to true differences in metacognitive bias given appropriate assumptions. Finally, applying it to a longitudinal dataset, we show that meta-Δ has high test-retest reliability, supporting its use in psychological research. Overall, our findings offer a principled framework for disentangling metacognitive sensitivity and bias from task-level performance, from response characteristics, and from each other.

Signal detection theoretic models of metacognition

Metacognition tasks combine two distinct tasks: participants must both discriminate between possible stimuli (the type 1 task) and determine whether their type 1 responses are correct (the type 2 task; Clarke et al., 1959). Classical signal detection theory (SDT) aims to characterize participants' performance for the type 1 task (Figure 1A; Green & Swets, 1966). Under SDT, the observer is presented with a noisy signal $x_1 \sim \mathcal{D}_S(d')$ following a distribution $\mathcal{D}$ dependent on a stimulus $S \in \{0, 1\}$ and the observer's sensitivity d'. Then, the observer must infer the value of S. Although other options are available (see, e.g., Meyer-Grant et al., 2025), it is common

Variable	Description
S	Stimulus value (0 or 1)
R	Type 1 response (0 or 1)
C	Confidence (type 2 response; 1 to K)
x_1	Evidence for type 1 decision
x_2	Evidence for type 2 decision
$N_{s,r,c}$	Number of trials with stimulus s, type 1 response r, and type 2 response c
$\mathbf{y}_s$	Trial counts for stimulus s
$\boldsymbol{\pi}_s$	Response probabilities for stimulus s
d'	Type 1 sensitivity to stimulus
c	Type 1 response criterion
meta-d'	Metacognitive (type 2) sensitivity
$M = \frac{\text{meta-}d'}{d'}$	Metacognitive efficiency
meta-c	Type 1 criterion for type 2 decision
meta-$c^r_{2,k}$	Type 2 criterion for response r and confidence level k
dmeta-$c^r_{2,k}$	Difference in type 2 criteria for response r between confidence levels $k-1$ and k
meta-Δ_r	Average distance between meta-$c^r_{2,k}$ and meta-c

Table 1: Summary of relevant variables, model parameters, and derived quantities of the meta-d' model

to assume $\mathcal{D}_0(d') = \mathcal{N}\left(-\frac{d'}{2}, 1\right)$ and $\mathcal{D}_1(d') = \mathcal{N}\left(\frac{d'}{2}, 1\right)$, where d' is the observer's sensitivity to the difference between the two stimuli. The observer makes a response by thresholding x by a criterion c (solid line in Figure 1A), such that the response is $R = [x_1 > c]$, and the observer makes a correct decision when $R = S$. This formulation is desirable because, in contrast to measures such as percent correct, it provides statistically independent measures of sensitivity (d') and response bias (c).

To measure metacognitive ability, Maniscalco and Lau (2012) developed an extension of SDT known as the meta-d' model (Table 1; Figure 1B-C; parameters for the type 2 decision beginning with "meta-" for clarity). Specifically, conditional on the type 1 response, the observer accesses a separate decision variable for making confidence ratings with sensitivity meta-d' bounded above or below by the type 1 criterion meta-c (bounds indicated with superscripts):

$$x_2 \sim \begin{cases} \mathcal{D}_S^{(-\infty,\text{meta-}c]}(\text{meta-}d') & \text{if } R = 0 \\ \mathcal{D}_S^{[\text{meta-}c,\infty)}(\text{meta-}d') & \text{if } R = 1 \end{cases}$$

Then, to determine the confidence level $C \in \{1 \dots K\}$, the observer rates confidence using a set of $K-1$ ordered confidence criteria (dashed lines in Figure 1B-C):

$$C = \begin{cases} 1 + \Sigma_{k=1}^{K-1}[x_2 < \text{meta-}c^0_{2,k}] & \text{if } R = 0 \\ 1 + \Sigma_{k=1}^{K-1}[x_2 > \text{meta-}c^1_{2,k}] & \text{if } R = 1 \end{cases}$$

Because of the parallel between the models for the type 1 and type 2 responses, the parameter meta-d' can be interpreted as the type 1 sensitivity of an ideal observer with the same response bias as the participant (Maniscalco & Lau, 2014). This interpretation allows for the direct comparison of d' and meta-d' in terms of the ratio $M = \frac{\text{meta-}d'}{d'}$, representing the observer's *metacognitive efficiency* (where $M = 1$ indicates optimal metacognition, $0 \leq M < 1$ indicates suboptimal metacognition, and $M > 1$ indicates super-optimal metacognition). It also allows for different approaches to equating

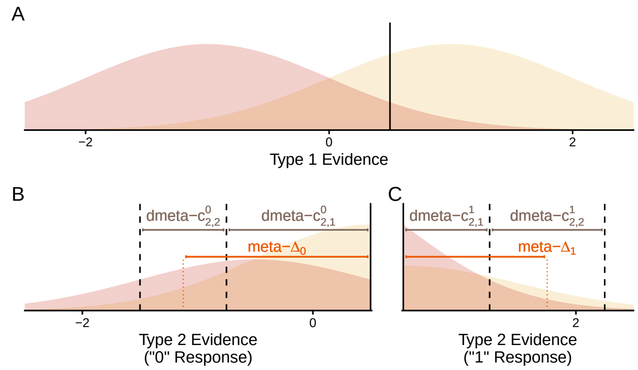


Figure 1: Overview of the meta-d' model and our measure of metacognitive bias, meta-Δ. (A) Evidence distributions underlying type 1 responses following one of two stimuli (yellow for $S = 1$, red for $S = 0$). The observer responds by thresholding the evidence according the response criterion (black line). (B-C) Response-specific distributions of evidence for type 2 decisions (confidence ratings). Confidence criteria (**meta-**$\mathbf{c_2^r}$; dashed lines) separate the amount of evidence required to report increasing levels of confidence. meta-Δ is distance between the type 1 criterion meta-c and the average of the confidence criteria (horizontal orange lines), which can be expressed as a weighted sum over the distances between subsequent confidence criteria (**dmeta-**$\mathbf{c_2^r}$; horizontal gray lines).

response bias for type 1 and type 2 responses. Here we focus on two possibilities: namely, meta-$c = c$ in the ABSOLUTE model (Fleming, 2017) and meta-$c = \frac{\text{meta-}d'}{d'}c$ in the RELATIVE model (Maniscalco & Lau, 2014).

Model parameterization We depict our formulation of the meta-d' model in Figure 2. Under this parameterization, all free parameters are derived from unconstrained variables to enable efficient optimization. In particular, we model the M-ratio on the logarithmic scale such that $M = e^{\text{log-}M}$. To ensure that confidence criteria are properly ordered, instead of modeling the confidence criteria (**meta-**$\mathbf{c_2^0}$ and **meta-**$\mathbf{c_2^1}$) directly, we model the differences between subsequent confidence criteria (**dmeta-**$\mathbf{c_2^0}$ and **dmeta-**$\mathbf{c_2^1}$). These differences are themselves constrained to be positive by modeling them on the logarithmic scale (**log-dmeta-**$\mathbf{c_2^0}$ and **log-dmeta-**$\mathbf{c_2^1}$).

Finally, for increased efficiency, we model the number of joint type 1 and type 2 responses for each value of the stimulus using a Multinomial likelihood. That is, we create an aggregated variable

$$\mathbf{y}_s = [N_{S=s,R=0,C=K}, \dots N_{S=s,R=0,C=1}, \\ N_{S=s,R=1,C=1}, \dots N_{S=s,R=1,C=K}]$$

where $N_{S=s,R=r,C=c}$ indicates the number of trials with type 1 response $R = r$ and type 2 response $C = c$ following the stimulus $S = s$. Then, the model estimates the corresponding response probabilities

$$\boldsymbol{\pi}_s = [P(R=0, C=K|S=s), \ldots P(R=0, C=1|S=s),$$
$$P(R=1, C=1|S=s), \ldots P(R=1, C=K|S=s)]$$

where $P(R = r, C = c \mid S = s)$ is the joint likelihood of a type 1 response $R = r$ and type 2 response $C = c$ for a trial where $S = s$, given by the meta-d' model as described above.

Quantifying metacognitive bias

The meta-d' model offers convenient measures of metacognitive sensitivity and efficiency that are relatively independent of type 1 sensitivity and bias (but see Guggenmos, 2021; Rausch et al., 2023; Xue et al., 2021). However, it remains unclear how best to quantify metacognitive bias in the context of this model. Sherman et al. (2018) developed a model-based measure of metacognitive bias which they denoted MDISTANCE, although found that this was also dependent on performance factors. Moreover, MDISTANCE is only defined for models with two confidence levels, limiting its applicability to designs with many confidence levels.

Here we present a new measure of metacognitive bias, meta-Δ, and demonstrate that, under the generative assumptions of the meta-d' model, it is independent of sensitivity (d'), response bias (c), and metacognitive sensitivity (meta-d'). Conceptually, meta-Δ is the distance between meta-c and the average of the confidence criteria:

$$\text{meta-}\Delta_r = \left| \frac{1}{K-1} \sum_{k=1}^{K-1} \left[\text{meta-}c^r_{2,k}\right] - \text{meta-}c \right|$$

Because summation is a linear operator, this is equivalent to the average of the distances between each confidence criteria and meta-c (orange lines in Figure 1B-C):

$$\text{meta-}\Delta_r = \frac{1}{K-1} \sum_{k=1}^{K-1} \left|\text{meta-}c^r_{2,k} - \text{meta-}c\right|$$

Then, since meta-$c^r_{2,k}$ = meta-$c + \sum_{j=1}^{k}$ dmeta-$c^r_{2,j}$ under the parameterization in Figure 2, one can express meta-Δ as a weighted sum of the distances between successive confidence criteria (gray lines in Figure 1B-C):

$$\text{meta-}\Delta_r = \sum_{k=1}^{K-1} \frac{K-k}{K-1} \text{dmeta-}c^r_{2,k}$$

For the case with two confidence levels (and a single confidence criterion per response), meta-Δ_0 = dmeta-$c^0_{2,1}$ and meta-Δ_1 = dmeta-$c^1_{2,1}$.

There are three notable features of meta-Δ. First, in contrast to MDISTANCE, we do not normalize meta-Δ by meta-d'. This is because doing so induces a spurious dependence of meta-Δ on meta-d'. Second, meta-Δ_r is calculated conditional on a particular response, allowing metacognitive bias to differ for "0" and "1" responses. If a single measure is preferable, we recommend taking the average meta-$\Delta = \frac{1}{2}\left[\text{meta-}\Delta_0 + \text{meta-}\Delta_1\right]$. As this average is unweighted, it reflects average metacognitive bias across the two responses, not the average metacognitive bias for a typical trial. If the latter is more desirable, one can alternatively compute meta-$\Delta' = P(R=0)\text{meta-}\Delta_0 + P(R=1)\text{meta-}\Delta_1$, but of course this version will be sensitive to differences in the relative proportions of each stimulus, as well as the participant's type 1 sensitivity and criterion. For example, a participant with high meta-Δ_0 and low meta-Δ_1 can systematically increase (decrease) meta-Δ' by decreasing (increasing) their type 1 criterion.

Third, meta-Δ is not strictly a measure of bias in the statistical sense because we do not compare it to an optimal value. Maniscalco et al. (2025) derived parametric formulae for the optimal confidence criteria under the meta-d' model for a range of different optimization targets (e.g., calibration between confidence and accuracy). But without access to the parameters underlying this optimization function (e.g., the desired accuracy for confidence level 2), one cannot directly compute optimal confidence criteria. Nevertheless, the scale of meta-Δ is meaningful in that greater values reflect a tendency for lower confidence ratings overall, with zero representing a participant who responds using only the highest confidence level and infinity representing a participant who responds using only the lowest confidence level.

Simulations

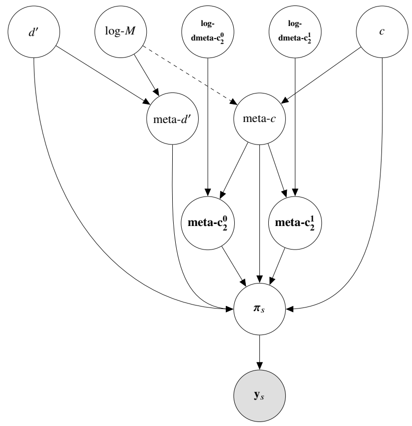
To confirm the validity of our measure, we ran three sets of simulations. In the first set of simulations, we manipulated the distances between the confidence criteria and meta-c (i.e., dmeta-$c^0_{2,1}$ and dmeta-$c^1_{2,1}$) to verify that meta-Δ is indeed sensitive to the amount of evidence required to make a response with high confidence. In the second set of simulations, we manipulated other model parameters (c, d', and meta-d') to confirm that meta-Δ is not confounded by these changes. In the final set of simulations, we performed the same manipulations but with data generated from a different generative model to test the extent to which meta-Δ is sensitive to assumptions about the generative process.

Methods

All models were fit using the `hmetad` package in R (Bürkner, 2017; O'Neill & Fleming, 2026). Parameters were independently manipulated using the values in Table 2, with all other parameters taking default values also shown in Table 2. After simulating datasets of 1,000,000 trials per parameter setting, we fit a single model to each dataset using the Laplace approximation with standard normal priors for all parameters. We then compared values of mean confidence and of meta-Δ conditional on each response implied by the fitted models. All simulation code is openly available at `https://github.com/kevingoneill/metacognitive_bias`.

Results

First, we confirmed that mean confidence and meta-Δ are sensitive to changes in the confidence criteria. Replicating results by Sherman et al. (2018), we found that mean confidence was



$$
\begin{aligned}
\mathbf{y}_s &\sim \text{Multinomial}(\boldsymbol{\pi}_s)\\
\boldsymbol{\pi}_s &= \text{metad\_pmf}_s(d', c, \text{meta-}d', \text{meta-}c, \textbf{meta-c}_\mathbf{2}^\mathbf{0}, \textbf{meta-c}_\mathbf{2}^\mathbf{1})\\
\text{meta-}d' &= \exp(\text{log-}M)\ d'\\
\text{meta-}c &= \begin{cases} c & \textsc{fixed} \\ \exp(\text{log-}M)\}\ c & \textsc{relative} \end{cases}\\
\textbf{meta-c}_\mathbf{2}^\mathbf{0} &= \text{meta-}c - \text{cumulative\_sum}(\exp(\textbf{log-dmeta-c}_\mathbf{2}^\mathbf{0}))\\
\textbf{meta-c}_\mathbf{2}^\mathbf{1} &= \text{meta-}c + \text{cumulative\_sum}(\exp(\textbf{log-dmeta-c}_\mathbf{2}^\mathbf{1}))
\end{aligned}
$$

Figure 2: Generative model for our parameterization of the meta-d' model (Maniscalco & Lau, 2012). Here, $\mathbf{y}_s$ and $\boldsymbol{\pi}_s$ are the joint type 1/type 2 response counts and response probabilities for each of the two stimuli, and metad_pmf_s is the likelihood of the meta-d' model for the stimulus $S = s$, log-M is the M-ratio on the logarithmic scale and $\textbf{log-dmeta-c}_\mathbf{2}^\mathbf{r}$ are the differences between successive confidence criteria for a type 1 response $R = r$ on the logarithmic scale.

Parameter	Default	Range	Step Size
dmeta-$c_{2,1}^0$	0.5	[0.1, 3.0]	0.1
dmeta-$c_{2,1}^1$	0.5	[0.1, 3.0]	0.1
c	0.5	[-1.5, 1.5]	0.1
d'	2	[0.5, 3.0]	0.1
meta-d'	1	[0.5, 3.0]	0.1

Table 2: Parameter values for model simulations. When manipulated, parameters were varied according to a grid determined by Range and Step Size. Otherwise, all parameters took their default values.

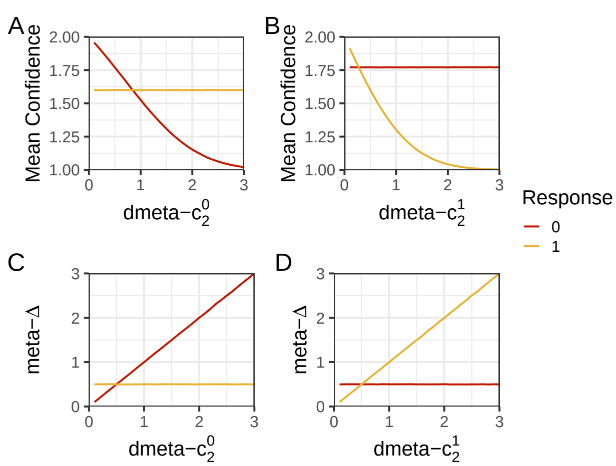


Figure 3: Estimated mean confidence and meta-Δ separately by response as a function of the confidence criteria meta-$c_{2,1}^0$ and meta-$c_{2,1}^1$ under proper model specification.

sensitive to changes in the confidence criteria. In particular, while the mean confidence of “0” responses decreased with increases in the confidence criterion for “0” responses, the mean confidence of “1” responses was unaffected by this change (Figure 3A). We found a similar effect when manipulating confidence criteria for “1” responses (Figure 3B). Likewise, response-specific meta-Δ increased only with the confidence criterion for its corresponding response (Figures 3C-D).

Next, we tested whether mean confidence and meta-Δ were robust to changes in conceptually unrelated model parameters. Again replicating results from Sherman et al. (2018), mean confidence had a response-specific and nonlinear relationship with response bias (c): although mean confidence in “0” responses increased with response bias, mean confidence in “1” responses showed the opposite trend (Figure 4A). We also found that mean confidence increased with sensitivity (d'; Figure 4B) and with metacognitive sensitivity (meta-d'; Figure 4C). In contrast, meta-Δ was unaffected by all three manipulations (Figure 4D-F).

Finally, to determine the extent to which each measure is sensitive to assumptions about the generative process of confidence ratings, we tested the behavior of mean confidence and meta-Δ under model mis-specification. That is, we generated data using the RELATIVE meta-d' model and fitted the ABSOLUTE model. As before, mean confidence was confounded by response bias, sensitivity, and metacognitive sensitivity in a non-linear and response-specific way (Figure 5A-C). However, in contrast to our earlier findings, we found that meta-Δ was also spuriously affected by changes in response bias, sensitivity, and metacognitive sensitivity when the generative assumptions of the model were violated (Figure 5D-F).

Application to real data

The above simulation studies reveal that, unlike mean confidence, meta-Δ provides a valid measure of metacognitive bias so long as assumptions about the generative process underly-

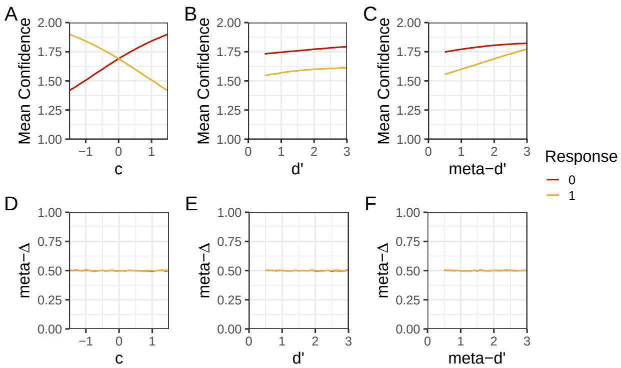


Figure 4: Estimated mean confidence and meta-Δ by response as a function of response bias (c), sensitivity (d'), and metacognitive sensitivity (meta-d') under proper model specification. Lines indicate posterior medians.

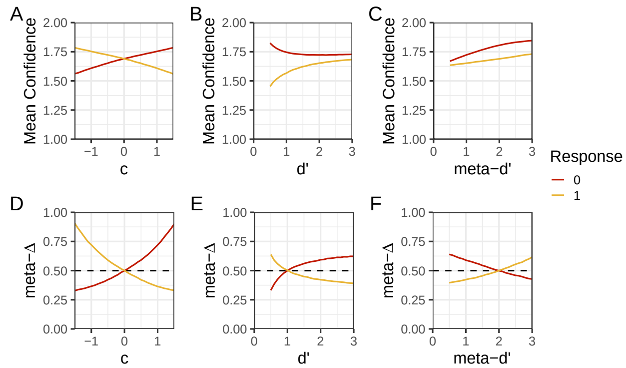


Figure 5: Estimated mean confidence and meta-Δ by response as a function of response bias (c), sensitivity (d'), and metacognitive sensitivity (meta-d') under model misspecification. Lines indicate posterior medians and dashed lines represent ground truth parameter values.

ing confidence data are met (see Rausch et al., 2023, for similar results concerning M-ratio). However, this measure is only useful for psychological research insofar as it is also reliable. Previous work has found that mean confidence is reliable (Ais et al., 2016; Fox et al., 2024; Rahnev, 2025) and predictive (Rouault et al., 2018; but see da Fonseca et al., 2023). To determine whether meta-Δ exhibits similar test-retest reliability, we applied this measure to a newly collected longitudinal dataset. All data and analysis code is openly available at `https://github.com/kevingoneill/metacognitive_bias`.

Methods

Using the `hmetad` package in R (O'Neill & Fleming, 2026), we applied the model to a longitudinal dataset obtained in a study examining links between confidence and information-seeking. Participants were recruited from Prolific (https://www.prolific.com), and the experiment was hosted on the online platform Gorilla. All protocols were approved by the University College London Research Ethics Committee (ID Number: EP_2024_004). Out of 250 recruited participants, 189 (75.6%) participants completed all four sessions on scheduled days with at least 90 valid trials in each session. We first removed trials where the reaction time of the first response exceeded 5 IQRs from the median (on average 1.02 ± 1.61 trials removed per session). Individual sessions were excluded if the accuracy of the type 1 response fell out of the [0.55, 0.90] range (N = 26 sessions, 3.44%), or there were three or more unused joint (type 1 x type 2) response levels (N = 321 sessions, 42.5%). Participants were further excluded if they failed to use four or more joint response levels (N = 11 participants, 5.82%). The final dataset consisted of 43,051 trials across 435 sessions from 157 participants (age: M = 31.77 , SD = 4.92 , range = 19 – 40 ; sex: 88 females, 67 males).

Participants were tested on four different days over one month using a perceptual decision-making task (Schulz et al., 2020). In the main task, participants were required to make a 2AFC decision about which side of the box contained more dots, report their confidence (3-level scale) in this decision, and then decide whether they prefer seeing the stimulus again before making the final decision. Opting to see the stimulus again costed task points, did not make the task longer, and would improve the final accuracy. Correct final decisions were rewarded. Participants were incentivized to maximize points across all sessions. Before the main task in each session, participants' performance were titrated to approximately 71% accuracy using a 2-down-1-up staircasing procedure.

To analyze the data, we fit a hierarchical version of the meta-d' model depicted in Figure 2 with session-level effects per participant for all parameters of the model. We used $\mathcal{N}(0, 0.25)$, $\mathcal{N}(1, 0.25)$, $\mathcal{N}(0, 0.1)$, and $\mathcal{N}(-.5, 1)$ priors over $\log(M)$, d', c, and $\log(\mathbf{dmeta\text{-}c_2})$, $\mathcal{N}^+(0, .5)$, $\mathcal{N}^+(0, .33)$, $\mathcal{N}^+(0, .33)$, and $\mathcal{N}^+(0, .5)$ priors for the standard deviations of participant-level differences in $\log(M)$, d', c, and $\log(\mathbf{dmeta\text{-}c_2})$, and LKJ(2) priors for all correlations between participant-level effects. We ran the model for 1000 iterations of warm-up and 1000 iterations of sampling for each of four chains of Hamiltonian Monte-Carlo. Although the fitted model had 115 divergent transitions, they were relatively infrequent (2.9% overall) and all other model diagnostics were acceptable. We calculated participant-level estimates of meta-Δ_0 and meta-Δ_1, as well as the average metacognitive bias meta-Δ, for each session from the fitted model. We report correlations between each of these estimates across sessions, as well as their 95% credible intervals.

Results

Despite significant posterior uncertainty for some participants, we found strong correlations as high as r = .77 [.69, .83] (meta-Δ, sessions 1 and 2) between participant-level estimates of meta-Δ_0, meta-Δ_1, and meta-Δ across the four sessions (Figure 6). The smallest correlation was r = .46 [.27, .61]

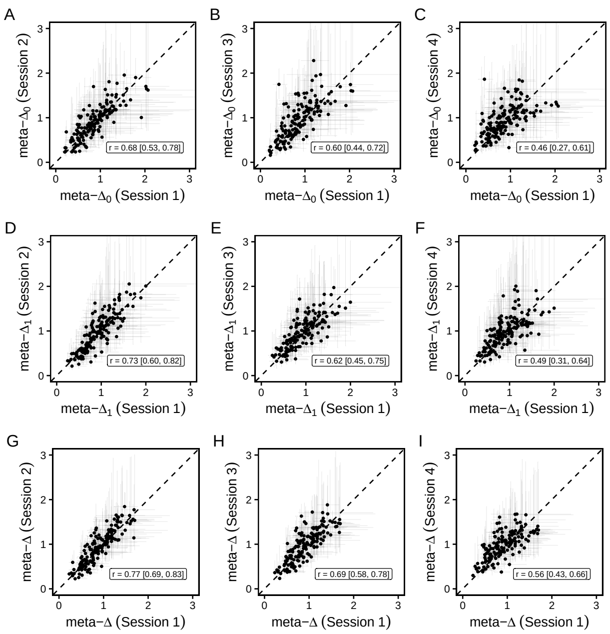


Figure 6: Estimated meta-Δ for (A-C) "0" responses, (D-F) "1" responses, and (G-I) averaging across responses, compared across four sessions of a perceptual 2AFC task. Points and errorbars are posterior medians and 95% credible intervals respectively.

(meta-Δ_0, sessions 1 and 4), indicating that meta-Δ exhibits strong test-retest reliability in perceptual tasks even across timescales as long as one month.

Discussion

Cognitive modeling often aims to separate different factors underlying performance: Signal Detection Theory, for example, distinguishes between an observer's sensitivity to a stimulus (d') from their overall response bias (c). Similarly, research on metacognition aims to quantify different ways in which confidence ratings relate to task performance. Significant developments have been made in terms of developing measures of metacognitive efficiency, that is the degree to which confidence ratings distinguish correct from incorrect responses (Fleming, 2017; Maniscalco & Lau, 2012). But despite ample evidence that metacognitive bias—the overall tendency to respond with high (or low) confidence—is clinically significant (Fox et al., 2024; Hoven et al., 2019; Katyal & Fleming, 2026; Katyal et al., 2025) and relevant to behaviors like information-seeking (Schulz et al., 2023) and evidence integration (Rollwage et al., 2020), less attention has been paid to developing validated measures. This raises the possibility that the existing literature paints a distorted picture of metacognitive bias.

Replicating results from Sherman et al. (2018), we found that the most frequently used measure of metacognitive bias, mean confidence, is confounded by sensitivity, response bias, and metacognitive sensitivity. Unfortunately, this means that unless experimentalists control for all relevant performance factors, differences in mean confidence cannot be expected to reflect true differences in metacognitive bias.

In response, here we developed a measure of metacognitive bias (meta-Δ) derived from a standard model of metacognitive sensitivity (the meta-d' model; Maniscalco & Lau, 2012). Our simulation results showed that, like mean confidence, meta-Δ is sensitive to true differences in the amount of evidence required to respond with high confidence. In contrast to mean confidence, we found that meta-Δ is independent of response bias, sensitivity, and metacognitive sensitivity. Finally, upon applying our measure to a recently collected longitudinal dataset, we found that meta-Δ has high test-retest reliability.

Our results provide strong evidence that, in the absence of strict methodological controls, meta-Δ has strictly better psychometric properties than mean confidence as a measure of metacognitive bias. But this success does not come without its limits; in line with earlier work showing that measures of metacognitive efficiency are sensitive to the generative process underlying confidence ratings (Guggenmos, 2021; Rausch et al., 2023), we found that meta-Δ is sensitive to the same assumptions. In particular, when the FIXED model was fitted to data generated by the RELATIVE model, we found that estimates of meta-Δ were biased in a non-linear and response-specific manner unless the true model had no response bias ($c = 0$) or theoretically optimal metacognition (meta-$d' = d'$). In such cases, neither mean confidence nor meta-Δ serve as valid measures of metacognitive bias. Overall, then, while meta-Δ serves as a useful measure of metacognitive bias when the generative assumptions underlying the meta-d' model are met, it is likely necessary for researchers to develop more domain-specific measures of metacognitive sensitivity and bias for use in other contexts.

Acknowledgments

SMF is a CIFAR Fellow in the Brain, Mind and Consciousness Program, and funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (selected as ERC Consolidator, grant number 101043666).

References

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, *146*, 377–386.

Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., Lau, J. Y., Latham, P. E., Bahrami, B., & Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*(6), 0117.

Bang, D., Ershadmanesh, S., Nili, H., & Fleming, S. M. (2020). Private–public mappings in human prefrontal cortex. *Elife*, *9*, e56477.

Boundy-Singer, Z. M., Ziemba, C. M., & Goris, R. L. (2023). Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour*, *7*(1), 142–154.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Clarke, F. R., Birdsall, T. G., & Tanner Jr, W. P. (1959). Two types of roc curves and definitions of parameters. *The Journal of the Acoustical Society of America*, *31*(5), 629–630.

da Fonseca, M., Maffei, G., Moreno-Bote, R., & Hyafil, A. (2023). Mood and implicit confidence independently fluctuate at different time scales. *Cognitive, Affective, & Behavioral Neuroscience*, *23*(1), 142–161.

Fleming, S. M. (2017). Hmeta-d: Hierarchical bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of consciousness*, *2017*(1), nix007.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience*, *8*, 443.

Fox, C. A., McDonogh, A., Donegan, K. R., Teckentrup, V., Crossen, R. J., Hanlon, A. K., Gallagher, E., Rouault, M., & Gillan, C. M. (2024). Reliable, rapid, and remote measurement of metacognitive bias. *Scientific Reports*, *14*(1), 14941.

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

Guggenmos, M. (2021). Measuring metacognitive performance: Type 1 performance dependence and test-retest reliability. *Neuroscience of consciousness*, *2021*(1), niab040.

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C. D., & Bahrami, B. (2017). Neural computations underpinning the strategic management of influence in advice giving. *Nature Communications*, *8*(1), 2191.

Hoven, M., Lebreton, M., Engelmann, J. B., Denys, D., Luigjes, J., & van Holst, R. J. (2019). Abnormalities of confidence in psychiatry: An overview and future perspectives. *Translational psychiatry*, *9*(1), 268.

Katyal, S., & Fleming, S. M. (2026). Gender and anxiety reveal distinct computational sources of underconfidence. *Psychological Medicine*, *56*.

Katyal, S., Huys, Q. J., Dolan, R. J., & Fleming, S. M. (2025). Distorted learning from local metacognition supports transdiagnostic underconfidence. *Nature Communications*, *16*(1), 1854.

Mamassian, P., & de Gardelle, V. (2025). The confidence-noise confidence-boost (cncb) model of confidence rating data. *PLOS Computational Biology*, *21*(4), e1012451.

Maniscalco, B., Charles, L., & Peters, M. A. (2025). Optimal metacognitive decision strategies in signal detection theory. *Psychonomic Bulletin & Review*, *32*(3), 1041–1069.

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. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta-d', response-specific meta-d', and the unequal variance sdt model. In *The cognitive neuroscience of metacognition* (pp. 25–66). Springer.

Meyer-Grant, C. G., Kellen, D., Harding, S. M., & Singmann, H. (2025). Extreme-value signal detection theory for recognition memory: The parametric road not taken.

O'Neill, K., & Fleming, S. (2026). *Hmetad: Fit the meta-d' model of confidence ratings using 'brms'* [R package version 0.1.0]. https://metacoglab.github.io/hmetad/

Rahnev, D. (2025). A comprehensive assessment of current methods for measuring metacognition. *Nature Communications*, *16*(1), 701.

Rausch, M., Hellmann, S., & Zehetleitner, M. (2023). Measures of metacognitive efficiency across cognitive models of decision confidence. *Psychological Methods*.

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature communications*, *11*(1), 2634.

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biological psychiatry*, *84*(6), 443–451.

Schulz, L., Fleming, S. M., & Dayan, P. (2023). Metacognitive computations for information search: Confidence in control. *Psychological Review*, *130*(3), 604.

Schulz, L., Rollwage, M., Dolan, R. J., & Fleming, S. M. (2020). Dogmatism manifests in lowered information search under uncertainty. *Proceedings of the National Academy of Sciences*, *117*(49), 31527–31534.

Seow, T. X., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How local and global metacognition shape mental health. *Biological psychiatry*, *90*(7), 436–446.

Sherman, M., Seth, A., & Barrett, A. (2018). Quantifying metacognitive thresholds using signal-detection theory. *bioRxiv*. https://doi.org/10.1101/361543

Van Marcke, H., & Desender, K. (2025). Context-dependent role of confidence in information-seeking. *Cognition*, *263*, 106219.

Xue, K., Shekhar, M., & Rahnev, D. (2021). Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Consciousness and Cognition*, *95*, 103196.