Measures of metacognitive efficiency across cognitive models of decision confidence

Manuel Rausch[1,2], Sebastian Hellmann[1], and Michael Zehetleitner[1]

[1] Catholic University of Eichstätt-Ingolstadt

[2] Rhine-Waal University of Applied Sciences

Author Note

Manuel Rausch https://orcid.org/0000-0002-5805-5544

Sebastian Hellmann https://orcid.org/0000-0001-6006-5103

Michael Zehetleitner https://orcid.org/0000-0003-3363-2680

This research was in part supported by grants ZE 887/8-1, ZE 887/9-1, RA2988/3-1 and RA2988/4-1 by the Deutsche Forschungsgemeinschaft. The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript. We have no conflicts of interest to disclose. All data and analysis code are available at https://osf.io/72uds. We are grateful to Veronika Eder for assistance in writing this paper.

Correspondence should be addressed to Manuel Rausch, Hochschule Rhein-Waal. Fakultät für Gesellschaft und Ökonomie, Marie-Curie-Straße 1, 47533 Kleve, Germany. E-Mail: manuel.rausch@hochschule-rhein-waal.de

Abstract

Meta-d′/d′ has become the quasi-gold standard to quantify metacognitive efficiency because meta-d′/d′ was developed to control for discrimination performance, discrimination criteria, and confidence criteria even without the assumption of a specific generative model underlying confidence judgments. Using simulations, we demonstrate that meta-d′/d′ is not free from assumptions about confidence models: Only when we simulated data using a generative model of confidence according to which the evidence underlying confidence judgements is sampled independently from the evidence utilized in the choice process from a truncated Gaussian distribution, meta-d′/d′ was unaffected by discrimination performance, discrimination task criteria, and confidence criteria. According to five alternative generative models of confidence, there exist at least some combination of parameters where meta-d′/d′ is affected by discrimination performance, discrimination criteria and confidence criteria. A simulation using empirically fitted parameter sets showed that the magnitude of the correlation between meta-d′/d′ and discrimination performance, discrimination task criteria, and confidence criteria depends heavily on the generative model and the specific parameter set and varies between negligibly small and very large. These simulations imply that a difference in meta-d′/d′ between conditions does not necessarily reflect a difference in metacognitive efficiency but might as well be caused by a difference in discrimination performance, discrimination task criterion, or confidence criteria.

Keywords: Metacognition, metacognitive efficiency, confidence, cognitive modelling, signal detection theory, meta-d′/d′

Metacognitive efficiency in cognitive models of decision confidence

A key aspect of metacognition is metacognitive efficiency, defined as a subject's level of metacognition given their discrimination task performance or signal processing capacity (Fleming & Lau, 2014). The gold standard to measure of metacognitive efficiency is meta-d′/d′ (Maniscalco & Lau, 2012, 2014). Measuring metacognitive efficiency by meta-d′/d′ has inspired research on many different psychological concepts, including learning (Boldt et al., 2019; Hainguerlot et al., 2018; Taouki et al., 2022), cognitive control (Drescher et al., 2018), vigilance (Maniscalco et al., 2017), memory (Mazancieux et al., 2020; Vandenbroucke et al., 2014), perception (Maniscalco et al., 2016; Odegaard, Chang, et al., 2018), psychopathology (Bhome et al., 2022; Culot et al., 2021; Muthesius et al., 2022; Rouault et al., 2018), beliefs about politicised science (Fischer & Said, 2021; Said et al., 2022), and visual awareness (Charles et al., 2013; Rausch & Zehetleitner, 2016; Vlassova et al., 2014). One reason why the meta-d′/d′ method has become so popular is that meta-d′ is believed to provide control over discrimination performance, discrimination task criteria, and confidence criteria (Maniscalco & Lau, 2012, 2014), which is a key requirement for measures of metacognitive accuracy (Barrett et al., 2013). Meta-d′ is also popular because it does not explicitly assume a specific generative model for confidence judgments (Maniscalco & Lau, 2014). However, there each exists at least one generative model of confidence which implies that meta-d'/d′ is affected by discrimination performance (Guggenmos, 2021) and confidence criteria (Shekhar & Rahnev, 2021), raising the question how robust meta-d′/d′ is with respect to the control over discrimination performance, discrimination task criteria, and confidence criteria across different generative models of confidence.

The meta-d′/d′ method

The meta-d′/d′ method is based on signal detection theory (Green & Swets, 1966; Peterson et al., 1954; Tanner & Swets, 1954) and type 2 signal detection theory (Clarke et al., 1959; Galvin et al., 2003; Pollack, 1959). The conceptual idea of meta-d′ is to quantify the accuracy of metacognition in terms of discrimination sensitivity in a hypothetical signal detection model describing the primary task, assuming participants had perfect access to the sensory evidence underlying the discrimination choice and were perfectly consistent in placing their confidence criteria (Maniscalco & Lau, 2012, 2014). Using a signal detection model describing the primary task to quantify metacognitive accuracy has the advantage of allowing a direct comparison between metacognitive accuracy and discrimination performance. Meta-d′ can be compared against the estimate of the distance between the two stimulus distributions estimated from discrimination responses, which is referred to as d′: If meta-d′ equals d′, it means that metacognitive accuracy is exactly as good as expected from discrimination performance. If meta-d′ is lower than d′, it means that metacognitive accuracy is worse than expected from discrimination performance (Fleming & Lau, 2014; Maniscalco & Lau, 2012, 2014).

The hypothetical signal detection model underlying meta-d′ assumes that the observer selects a binary response $R \in \{-1, 1\}$ about a stimulus characterised by two classes $S \in \{-1, 1\}$ as well as a confidence rating out of an ordered set of confidence categories $C \in \{1, 2, \ldots, n\}$ (see Table 1 for a list of our mathematical notation). For each presentation of the stimulus, the observer's perceptual system creates sensory evidence delineating the two response options. As there is noise in the system, the sensory evidence is not constant, but modelled as a random sample x out of a separate Gaussian distribution for each of the two stimulus classes (see Fig. 1). The distance d between the two distributions created by the two classes of S is interpreted as the observer's ability to differentiate between the two kinds of S. Participants select a response by comparing the sensory evidence x with a response criterion c, choosing R = -1 if the sensory evidence x is smaller than the response criterion, and R = 1 otherwise. Confidence ratings are chosen by comparing the same sample of sensory evidence x against a set of $2 \times n - 1$ confidence criteria, θ_1, θ_2, $\theta_3, \ldots,$ $\theta_{2\times n-1}$. For example, if there are four confidence categories, participants are assumed to select a response R of 1 and a confidence level of 3 if the sensory evidence x is smaller than the outermost response criterion θ_7, but at the same time greater than the second outermost response criterion θ_6.

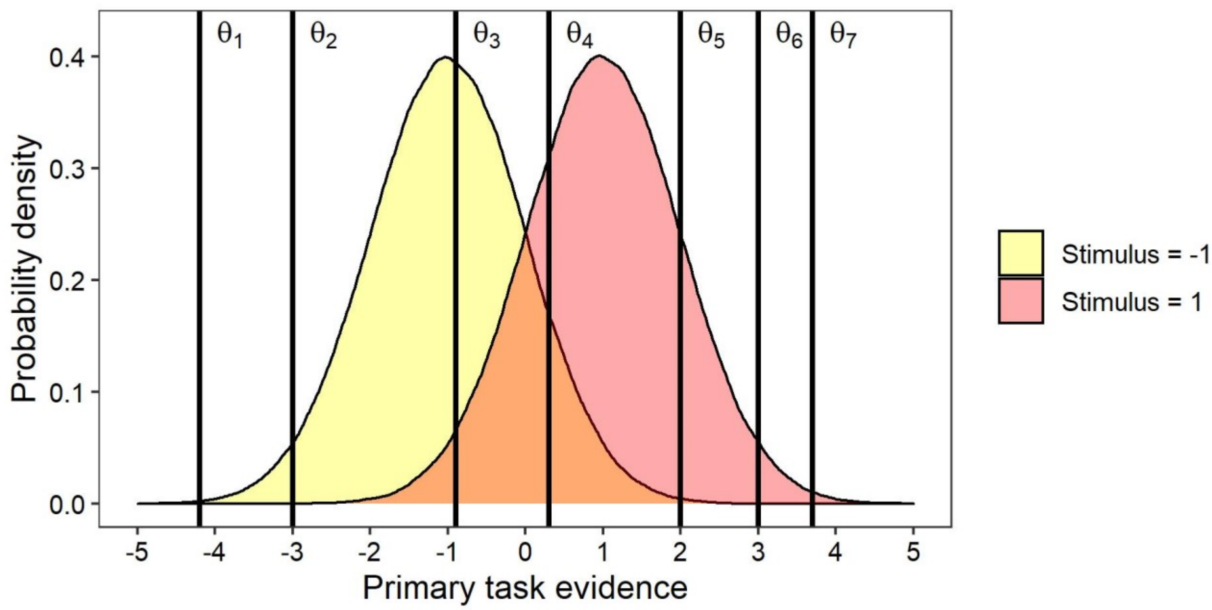
Table 1

Table of mathematical notation and terminology

Symbol	Description or terminology
S	Stimulus class
R	Discrimination response about the stimulus class
C	Confidence judgment
n	Number of options given by the confidence scale
x	Sensory evidence about S
d	distance between the two distributions of evidence created by the two different stimulus classes, interpreted as the observer's ability to differentiate between the two stimulus classes
d'	Estimate of d based on R
d_{meta}	Meta-d′: Estimate of d based on C
c	Response criterion for the discrimination judgment
θ	Criterion for confidence judgments
m	Metacognitive efficiency parameter within the independent truncated Gaussian model
y	Confidence decision variable

Figure 1

The hypothetical signal detection theoretic model underlying meta-d′



Note. The hypothetical signal detection theoretic model describing the primary task underlying meta-d′ (Maniscalco & Lau, 2012, 2014). To estimate meta-d′, it is assumed that the same evidence is available for selecting a response for the discrimination task and for selecting a confidence judgement. Primary task responses and confidence categories are assumed to form an ordered set of responses delineated by a set of criteria θ.

Meta-d′ vs. generative models of confidence

According to Maniscalco and Lau (2014), the meta-d′/d′ method only makes assumptions about the cognitive architecture underlying the discrimination choice, but meta-d′/d′ does not require an *explicit* assumption about the generative model underlying confidence judgments. However, it should be noted that the hypothetical signal detection model underlying meta-d′ is not dissimilar to the approach taken in studies that aim to identify the generative model underlying confidence judgments. The reason is that the estimation methods available to fit meta–d′ require the computation of the probability of the different levels of confidence given stimulus and discrimination response $p(C|R,S)$. Notably, static generative models of confidence are usually defined by a probability density of confidence ratings and discrimination task responses $p(C, R|S)$ (e.g. Adler & Ma, 2018; Aitchison et al., 2015; Rausch et al., 2018, 2020; Shekhar & Rahnev, 2021). This means what distinguishes the meta-d′ approach from generative models of confidence is whether the probability density is conditioned on the discrimination response or whether the discrimination response is modelled as well. According to both the conditioned maximum likelihood procedure proposed by Maniscalco and Lau (2014) and the Bayesian Markov Chain Monte Carlo (MCMC) method by Fleming (2017), the probability for a specific degree of confidence given stimulus and response $p(C|R, S)$ is given by

$$p(C = i|S, R = -1) = \frac{\int_{\theta_{n-i}}^{\theta_{n-i+1}} \phi_{\mu=d_{meta}\times S\times 0.5}(y)\, dy}{\int_{-\infty}^{\theta_n} \phi_{\mu=d_{meta}\times S\times 0.5}(y)\, dy} \tag{1}$$

$$p(C = i|S,\ R = 1) = \frac{\int_{\theta_{n+i-1}}^{\theta_{n+i}} \phi_{\mu=d_{meta}\times S\times 0.5}(y)\, dy}{\int_{\theta_n}^{\infty} \phi_{\mu=d_{meta}\times S\times 0.5}(y)\, dy} \tag{2}$$

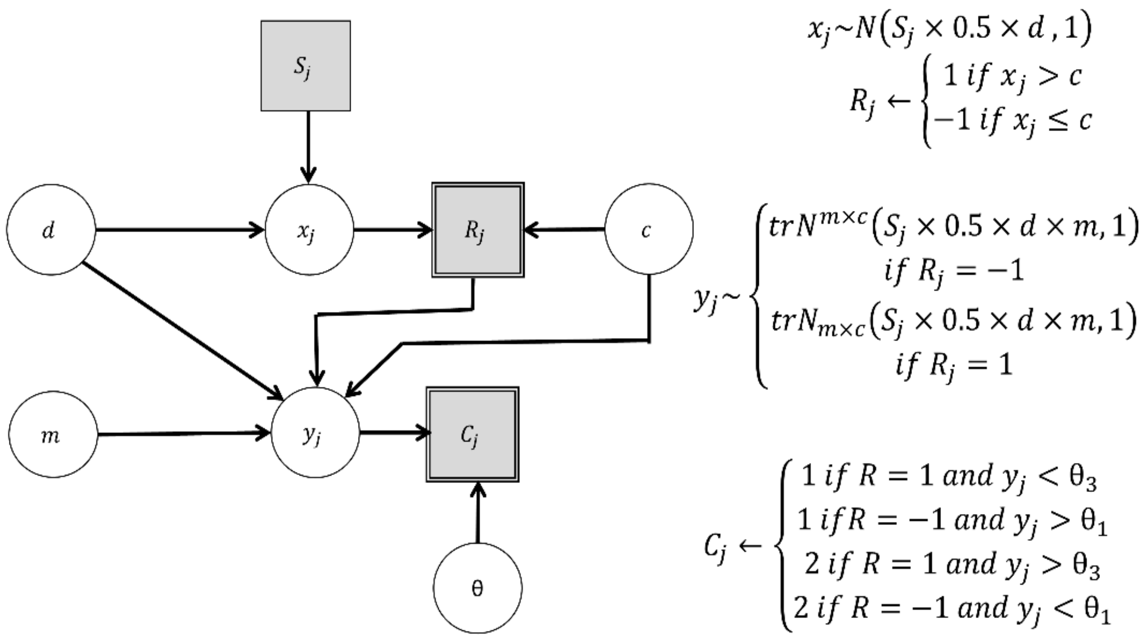
where ϕ indicates the Gaussian density function with mean μ and variance of 1, θ_0 is -∞, θ_{2n} is ∞, and d_{meta} is meta-d′. According to Maniscalco and Lau (2014), the location of the central confidence criterion θ_n depends on the perceptual sensitivity of the observer d′ as well as on the primary task criterion c and is given by $\theta_n = c \times d_{meta} \div d'$. According to Fleming's method, θ_n is identical to c. The formulae (1) and (2) show two important features of the meta-d′/d′ method. First, the formulae for $p(C|S, R)$ are identical to the cumulative truncated gaussian distribution function (Kristensen et al., 2020). Second, the formulae do not include x, the sensory evidence used to make the discrimination choice: This means that the random process underlying confidence judgments only depends on the outcome of the random process underlying the discrimination task decision, i.e., the response R, but when conditioned on R, it does not depend on the state of the random process generating the discrimination task decision.

The independent truncated Gaussian model (ITG)

Here, we present a generative model of confidence that is set up to be consistent with the probability functions used to estimate meta-d′: the *independent truncated Gaussian model* (ITG, see Fig. 2). Conceptually, ITG reflects a cognitive mechanism where confidence judgments are based on information generated independently from the sensory evidence used to make the perceptual decision. However, according to ITG, confidence judgments can only be informed by information corroborating the perceptual decision; contradicting information is not available. ITG is identical to standard signal detection theory as far as the discrimination task response is concerned. For the choice about the confidence, according to ITG, there is a separate decision variable for confidence y. The confidence decision variable y is sampled from a truncated Gaussian distribution, with the location parameter equal to $S \times d \times 0.5 \times m$ and a scale parameter of 1. The parameter d quantifies the perceptual ability of the observer and is equivalent to d′ in standard signal detection theory. The parameter m quantifies metacognitive efficiency, which is measured by meta-d′/d′. Notably, y is sampled independently from x, the sensory evidence used in the discrimination decision (see Fig. 3 for a visualisation of the distribution of x and y). The Gaussian distribution of y is truncated in a way that it is impossible to sample evidence that contradicts the original decision: If R = -1, the distribution is truncated to the right of θ_n. If R = 1, the distribution is truncated to the left of θ_n. Because Maniscalco and Lau (2014) and Fleming (2017) defined θ_n differently, there are also two slightly different versions of ITG. ITG reproduces the probability density of confidence given stimulus and response specified by Maniscalco and Lau (2014) if the distribution of y is truncated at $c \times m$, while to reproduce the probability density of confidence given stimulus and response in Fleming (2017), the distribution must be truncated at c. Just as in the signal detection model, confidence ratings are chosen by comparing the confidence decision variable y against a set of $2 \times n - 1$ confidence criteria, $\theta_1, \theta_2, \theta_3, \ldots, \theta_{2\times n-1}$.

Figure 2

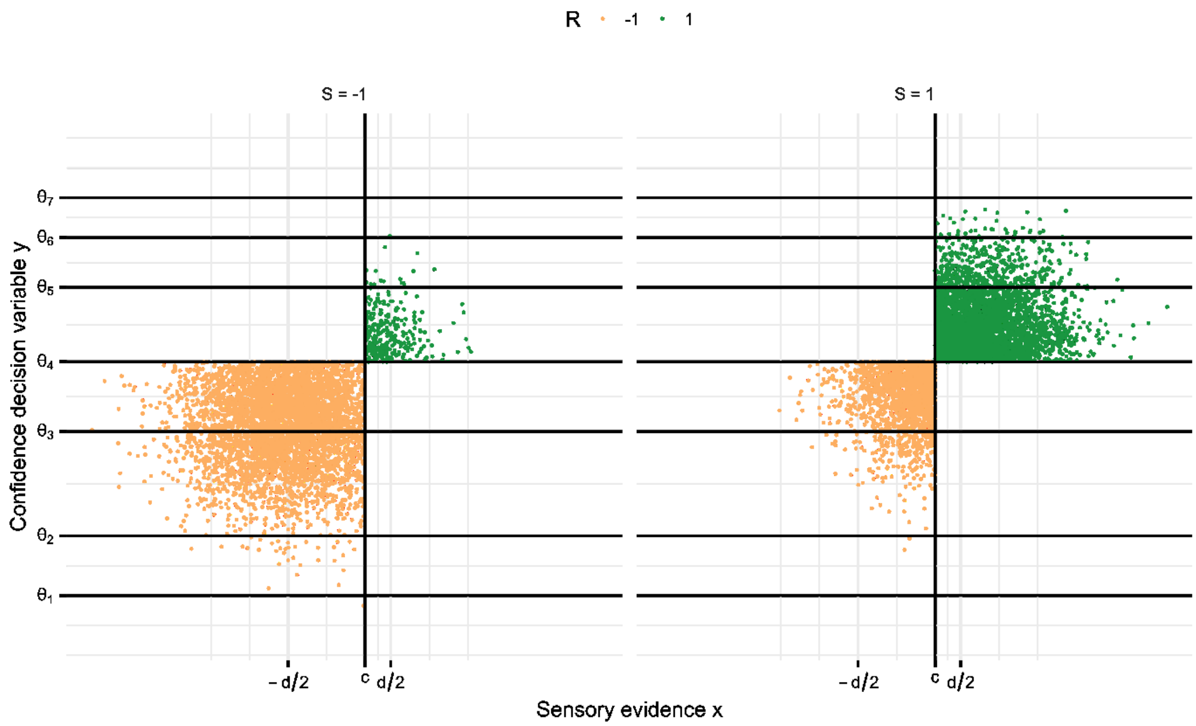
Bayesian graphical model of the independent truncated Gaussian model (ITG)



Note. Version of ITG to reproduce the probabilities of confidence categories given stimulus and response underlying the maximum likelihood method devised by Maniscalco and Lau (2014). S_j,, R_j, and C_j are stimulus class, response, and confidence in trial j, respectively , d is the discrimination sensitivity parameter, c is the discrimination criterion, θ is the confidence criterion, m is the metacognitive efficiency parameter, x_j is the sensory evidence in trial j, and y_j is the confidence decision variable in trial j. trN_a^b indicates a Gaussian distribution which is truncated at the left side and at b at the right side. Following the convention by Lee and Wagenmakers (2013), continuous variables are depicted as circles and discrete variables as squares, observed variables are shaded, unobserved variables not shaded, stochastic dependence is indexed by single borders, and deterministic dependence by double borders.

Figure 3

Two-dimensional distributions of sensory evidence x and confidence decision variable y according to the independent truncated Gaussian model (ITG)



Note. Fig. 3 is based on a simulation of the ITG model, using Fleming’s model specification, and assuming the following parameters: d = 2, c = 0.5, m = 0.5.

The implications of the similarity of the meta-d′ method and the ITG model with respect to the interpretation of meta-d′/d′ has to our knowledge not yet been explored: In standard signal detection theory, measures of sensitivity are only guaranteed to be independent from response criteria if the underlying SDT model is a reasonable approximation of the underlying processes (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002). Unfortunately, examples of generative models have been presented where meta-d′ is not robust against a variation of discrimination task performance and confidence criteria: According to a model where the confidence criteria are affected by lognormal noise, meta-d′/d′ is influenced by confidence criteria (Shekhar & Rahnev, 2021). According to a Bayesian model where confidence is affected by beta-distributed metacognitive noise, meta-d′/d′ depends on discrimination task performance (Guggenmos, 2021). Thus, the question arises how robust the control that meta-d′/d′ provides over discrimination task performance, discrimination task criterion, and confidence criteria is if the space of different generative models underlying confidence is varied more widely.

Rationale of the present study

In the present study, we investigated whether meta-d′/d′ is influenced by discrimination task performance, discrimination task criterion, and confidence criteria. For this purpose, we simulated artificial data while systematically varying the underlying generative model of confidence. Because the number of generative models of confidence proposed in the literature is far greater than what can be investigated in a single study (e.g. Desender et al., 2021; Fleming & Daw, 2017; Guggenmos, 2022; Mamassian & de Gardelle, 2021; Rausch et al., 2018; Maniscalco & Lau, 2016; Shekhar & Rahnev, 2021; Reynolds et al., 2020; Hellmann et al., 2023; Boundy-Singer et al., 2022; Zhu et al., 2023; Moran et al., 2015; Pereira et al., 2021), for the purpose of the present study, we restricted our analysis to models where the discrimination task decision is made consistent with signal detection theory and thus applying a meta-d′/d′ model is considered appropriate (Fleming & Lau, 2014). Besides two versions of the independent truncated gaussian model, one equivalent to the hypothetical SDT models used by Maniscalco and Lau (2014) and one equivalent to the hypothetical SDT models used by Fleming (2017), we used five different models reflecting different cognitive mechanisms how confidence judgments may be generated (see Table 2). For each simulation, we computed meta-d′/d′ using three different methods: 1) the conditioned maximum likelihood method proposed by Maniscalco and Lau (2012, 2014), 2) the Bayesian MCMC method described by Fleming (2017), and 3) conditioned maximum likelihood estimation using Fleming's specification of the hypothetical SDT model.

Table 2

List of cognitive models in which we analyzed the behavior of meta-d′/d′

Model	Reference	Conceptual interpretation of the model
Independent truncated Gaussian model	Maniscalco and Lau (2014) Fleming (2017)	Information used for confidence is generated independently from the evidence used for the choice. Evidence contradicting the original choice cannot be collected.
Postdecisional accumulation model	Pleskac and Busemeyer (2010)	After the choice, accumulation of sensory evidence continues for a fixed time interval
Gaussian noise model	Maniscalco and Lau (2016)	Confidence is informed by the same sensory evidence as the task decision, but confidence is affected by additive Gaussian noise.
Response-congruent evidence model	Maniscalco et al. (2016) Peters et al. (2017)	Confidence is informed only by evidence supporting the selected decision option; evidence in favor of the other option is ignored
Confidence boost model	Mamassian and de Gardelle (2021)	Confidence is informed by the evidence used for the choice and by evidence collected in parallel to the choice. In addition, confidence is affected by additive Gaussian noise.
Weighted evidence and visibility model	Rausch et al. (2018, 2020, 2021)	Confidence is informed by the evidence used for the choice as well as by a parallel estimate of the difficulty of the task. In addition, confidence is affected by additive Gaussian noise.

We expected that meta-d/d′ is independent from discrimination task performance, discrimination task criteria, and confidence criteria when the generative model is the independent truncated Gaussian model. At least for some of the alternative models, we expected that meta-d′/d′ depends on discrimination task performance, discrimination task criterion, and confidence criteria.

Simulation 1

Method

Model specification

We simulated data using seven different generative models:

i. the independent truncated Gaussian model with the Gaussian distribution truncated at the discrimination task criterion multiplied with metacognitive efficiency (consistent with the hypothetical SDT model proposed by Maniscalco and Lau, 2014),
ii. the independent truncated Gaussian model with the Gaussian distributions truncated at the discrimination task criterion (consistent with the hypothetical SDT model used by Fleming (2017),
iii. the Gaussian noise model,
iv. the postdecisional accumulation model,
v. the weighted evidence and visibility model,
vi. the confidence boost model, and
vii. the response-congruent evidence model.

For all seven models, we assumed that participants select a discrimination response $R \in \{-1, 1\}$ about the stimulus class $S \in \{-1, 1\}$ as well as a confidence judgment on a five-point scale that the response about the stimulus is correct $C \in \{1, 2, 3, 3, 5\}$. According to all seven models, a decision about the stimulus is made by comparing the sensory evidence x against the decision criterion c. Participants respond R = -1 if x < c and R = 1 if x > c. The sensory evidence x is modelled as a random sample from a Gaussian distribution:

$$x \sim N(\mu = S \times 0.5 \times d, \sigma = 1)$$

The more sensitive the observer is to the stimulus, the greater is the distance d between the centres of the distributions created by the two stimuli. Thus, d is interpreted as the ability of the observer's perceptual system to differentiate between the two kinds of S. The different models are characterised by ways how the confidence decision variable y is generated. A specific degree of confidence is determined by comparing y against a set of confidence criteria. To be consistent with standard SDT, we assumed separate of confidence criteria for each of the two response options. For all models, we assumed for simplicity that confidence criteria are placed symmetrically around the central confidence criterion θ_5 with the placement of criteria determined by the parameter τ. For the version of ITG modelled after Maniscalco and Lau's method, θ_5 was set to $c \times m$. For the version of ITG modelled after Fleming's method, as well as for the five alternative models of confidence, θ_5 was set to c. For R = -1, the other confidence criteria are located at $\theta_1 = \theta_5 - 2 \times \tau$, $\theta_2 = \theta_5 - 1.5 \times \tau$, $\theta_3 = \theta_5 - \tau$, and $\theta_4 = \theta_5 - 0.5 \times \tau$. For R = 1, the confidence criteria are located at $\theta_6 = \theta_5 + 0.5 \times \tau$, $\theta_7 = \theta_5 + \tau$, $\theta_8 = \theta_5 + 1.5 \times \tau$, and $\theta_9 = \theta_5 + 2 \times \tau$. Each criterion delineates between two adjacent confidence criteria, e.g., the observer reports confidence C = 2 if the response R is -1 and y fell between θ_1 and θ_2, or if R = 1 and y fell between θ_6 and θ_7. Thus, τ represents how liberally or conservatively participants place their confidence criteria.

Gaussian noise model. Conceptually, the Gaussian noise model reflects the idea that confidence is informed by the same sensory evidence as the task decision, but confidence is affected by additive Gaussian noise. Therefore, the confidence decision variable y is also sampled from a Gaussian distribution, with a mean equal to the sensory evidence x and a standard deviation σ_c, an additional free parameter.

$$y \sim N(\mu = x, \sigma = \sigma_c)$$

Postdecisional accumulation model. The postdecisional accumulation model was inspired by two-stage signal detection theory, according to which accumulation of sensory evidence is continued after the decision for a fixed time interval (Pleskac & Busemeyer, 2010). To ensure comparability with the other models, we used a model that represents the conceptual idea of ongoing accumulation of evidence but does not model reaction time data as well. According to PDA, the confidence decision variable y is sampled from a Gaussian distribution:

$$y \sim N(\mu = x + S \times 0.5 \times d \times b, \sigma = \sqrt{b})$$

The free parameter b indicates the amount of postdecisional accumulation relative to the amount of evidence available at the time of the discrimination decision. The standard deviation equals the square root of b because both the mean and the variance of the decision variable increase linearly with time in drift diffusion processes (Pleskac & Busemeyer, 2010).

Weighted evidence and visibility model. The conceptual idea underlying the weighted evidence and visibility model is that the observer combines evidence about the choice-relevant feature of the stimulus with the strength of evidence about choice-irrelevant features to select one out of several confidence categories (Rausch et al., 2018, 2020, 2021). Evidence about choice-irrelevant features of the stimulus can improve confidence judgement because they allow the observer to estimate the reliability of the percept more precisely (Rausch & Zehetleitner, 2019). To express this idea in formal terms, the WEV model assumes that y is sampled from a Gaussian distribution with the standard deviation σ_c:

$$y \sim N(\mu = (1 - w) \times x + w \times d \times R, \sigma = \sigma_c)$$

The standard deviation σ_c quantifies the amount of unsystematic variability contributing to confidence judgments but not to identification judgments. The unsystematic variability may stem from different sources, including the uncertainty in the estimate of stimulus strength or the noise inherent to metacognitive processes. The factor R ensures that strong stimuli tend to shift the location of the distribution in a way that high confidence is more likely, and likewise, weak stimuli tend to shift the location of the distribution in a way that the probability of low confidence increases.

Confidence boost model. The confidence boost model represents the idea that the confidence decision variable y is only partially based on the information used during the perceptual decision (Mamassian & de Gardelle, 2021). The confidence boost reflects information used for confidence judgments which was not used for perceptual decision. For this purpose, the model includes the parameter α, which quantifies the degree to which observer base their confidence judgments on information available for the perceptual decision. If $\alpha = 0$, confidence judgments are exclusively based on information already used for the perceptual decisions; if $\alpha = 1$, the observer has direct access to the original stimulus, and not just the noisy sensory evidence used to make the perceptual decision. In addition, there is again confidence noise superimposed on the confidence decision variable σ_c. Because Mamassian and de Gardelle (2021) conceived their model for confidence forced choice paradigms, the model was slightly adapted to be applicable for tasks where meta-d′/d′ is typically used. In the version of the model used in the present study, y is sampled from a Gaussian distribution with the standard deviation σ_c:

$$y \sim N(\mu = 0.5 \times S \times d + x \times (1 - \alpha), \sigma = \sigma_c)$$

Response-congruent evidence model. The model was inspired by the confidence model proposed by Peters et al. (2017). Conceptually, the model represents the idea that observers use all available sensory information to make the primary task decision, but for confidence judgments, they only consider evidence consistent with the selected decision and ignore evidence against the decision (Maniscalco et al., 2016; Odegaard, Grimaldi, et al., 2018; Samaha et al., 2016; Zylberberg et al., 2012). In our version of the model, the response-congruent evidence model assumes two separate samples of sensory evidence collected in each trial, each belonging to one possible identity of the stimulus:

$$x_1 \sim N(\mu = \ (1 - S) \ \times 0.25 \times d, \sigma = \sqrt{1/2})$$

$$x_2 \sim N(\mu = \ (1 + S) \ \times 0.25 \times d, \sigma = \sqrt{1/2})$$

The sensory evidence used for the discrimination choice is $x = x_2 - x_1$, which implies that the discrimination decision is equivalent to standard signal detection theory. The confidence decision variable depends on the response selected by the observer:

$$y = \begin{cases} -x_1, if \ \ R = -1 \\ \ \ x_2, if \ \ R = 1 \end{cases}$$

Simulations

Table 3 lists all parameters we used for our simulations. The parameters were chosen to investigate the behaviour of meta-d′/d′ across a decent range while at the same time avoiding extreme frequencies of events, which are known to lead to unstable behaviour (Barrett et al., 2013). For each generative model, we performed one simulation for each possible combination of parameters. In each simulation, we randomly simulated 10^6 discrimination responses and confidence ratings for both stimuli. Then, we computed meta-d′/d′ using three different methods:

i. the conditioned maximum-likelihood method as described by Maniscalco and Lau (2014),
ii. the Bayesian MCMC method used by Fleming (2017),

iii. a conditioned maximum-likelihood method that uses the specification of the hypothetical SDT model used by Fleming (2017).

A simulation was only included into the results if the estimated standard error of meta-d′ was below .005. All analyses were conducted using R (R Core Team, 2020).

Table 3

Parameters for each generative model of confidence

Model	Parameter	values used during simulations	Interpretation of the parameter
All models	d	0.5, 1.0, 1.5, 2.0, 2.5	sensitivity of the observer to discriminate between the two stimulus classes
	c	0, 0.25, 0.5, 1, 1.5, 2	criterion for the primary task response
	τ	0.5, 1.0, 1.5, 2.0, 2.5	placement of confidence criteria
Independent truncated Gaussian model	m	0.5, 1, 1.5	Amount of signal available for metacognition relative to the signal available for the discrimination choice
Gaussian noise model	σ_c	0.5, 1, 2	amount of noise superimposed on rating response
Postdecisional accumulation model	b	0.1, 0.5, 1	amount of postdecisional accumulation relative to the evidence available at the time of the discrimination decision
Weighted evidence and visibility model	σ_c	0.5, 2	amount of Gaussian noise superimposed on rating response
	w	0.25, 0.75	degree to which confidence relies on sensory evidence about the identity or on strength of evidence about identification-irrelevant features of the stimulus
Confidence boost model	σ_c	0.5, 2	amount of normal noise superimposed on rating response
	α	0.25, 0.75	degree to which observer has direct access to the original stimulus when making the confidence judgment

Conditioned maximum likelihood estimation of Maniscalco and Lau's model. To estimate meta-d′ based on conditioned maximum likelihood estimation, we used a translation of the MATLAB code provided by Brian Maniscalco (http://www.columbia.edu/~bsm2105/type2sdt, last accessed 2021-09-20) to R. The algorithm involved the following computational steps: First, the frequency of each confidence category was determined depending on the stimulus class and the accuracy of the response. To correct for extreme proportions, $1/(2n)$ was added to each cell of the frequency table. Second, discrimination sensitivity d' and discrimination criterion c were calculated using standard formulae

$$d' = \Phi^{-1}(\frac{n_{S1R1}}{n_{S1}}) - \Phi^{-1}(\frac{n_{S0R1}}{n_{S0}}) \tag{3}$$

$$c = -\frac{1}{2} \times \left(\Phi^{-1}(\frac{n_{S1R1}}{n_{S1}}) + \Phi^{-1}(\frac{n_{S0R1}}{n_{S0}})\right) \tag{4}$$

with n_{S1} the number of trials when $S = 1$, n_{S0} the number of trials when $S = -1$, n_{S1R1} the number of trials when $S = 1$ and $R = 1$, n_{S0} the number of trials when $S = -1$, n_{S0R1} the number of trials when $S = -1$ and $R = 1$, and Φ^{-1} the quantile function of the standard Gaussian distribution. The third step involved fitting the meta-d′ model. For this purpose, a maximum likelihood optimization procedure was used with respect to the probability of confidence given stimulus and response as well as the parameters determined at previous steps, i.e., d' and c. Model fitting involved a free parameter for meta-d′ d_{meta} as well as the rating criteria θ_1, θ_2, …, θ_{n-1}, θ_{n+1}, θ_{n+2}, …, θ_{2n-1}. To reproduce the original method by Maniscalco and Lau, θ_n was fixed at $c \times d_{meta} \div d'$. To enforce that the criteria were ordered, all free criteria were parametrized as the log of the distance to the adjacent criterion. Model fitting was performed in two steps: First, a coarse grid search was used to identify promising starting values. Second, the five best parameter sets were used as initial values for an Nelder-Mead optimization algorithm as implemented in the R function optim (Nelder & Mead, 1965). We restarted the optimization four times, using the previously found result as initial value for the next iteration to prevent the algorithm from getting stuck in a local minimum. Standard errors associated with the estimate of meta-d′ were obtained by inverting the Hessian matrix returned from optim.

Conditioned maximum likelihood estimation of Fleming's model. To fit meta-d′/d′ using conditioned maximum likelihood estimation and a model specification equivalent to the method used by Fleming (2017), we used the same algorithm as for Maniscalco and Lau's model specification with the exception that θ_n was fixed at c.

Bayesian Markov Chain Monte Carlo. To estimate meta-d′/d′ using Bayesian MCMC, we used R code provided by Steve Fleming (https://github.com/metacoglab/HMeta-d, last accessed 2022-10-22), which relies on the free software jags to sample from the posterior distribution (Plummer, 2003). For more details on the underlying Bayesian estimation procedure, see Fleming (2017). Just as for standard meta-d′, discrimination performance d′ and discrimination criterion c were computed first using formulae (3) and (4) and then submitted to jags as constants. The Bayesian estimation procedure was used only for the meta-d′/d′ and confidence criteria. For this purpose, the absolute frequency of each confidence rating given stimulus and response $f(C|S,R)$ was modelled as a multinomial distribution $\mathcal{M}$,

$$f(C|S,R) \sim \mathcal{M}(n = n_{SR}, p = p(C|S,R)) \tag{5}$$

where n_{SR} is the number of trials with stimulus S and response R, and $p(C|S,R)$ calculated using formulae (1) and (2). θ_n was fixed at c. $p(C|S,R)$ depends on the free parameters d_{meta} and a set of criteria θ. The priors for the parameters were specified as follows:

$$\theta_{1,2,\ldots,n-1} \sim tr\mathcal{N}\left(\mu = 0, \quad \sigma = \sqrt{0.5}, \quad a = -\infty, \quad b = c\right) \tag{6}$$

$$\theta_{n+1,n+2,\ldots,2\times n-1} \sim tr\mathcal{N}\left(\mu = 0, \quad \sigma = \sqrt{0.5}, \quad a = c, \quad b = \infty\right)$$

$$d_{meta} \sim \mathcal{N}(\mu = d', \sigma = \sqrt{2})$$

where $\theta_{1,2,\ldots,n-1}$ indicates the set of confidence criteria when the response was -1, $\theta_{n+1,n+2,\ldots,2\times n-1}$ indicates the set of confidence criteria when the response was 1, $tr\mathcal{N}$ indicates a truncated gaussian distribution with a location parameter μ, scale parameter σ, lower bound a, and upper bound b, and d_{meta} is meta-d′. These priors reflect the standard settings. Sampling was performed in three separate Markov Chains to allow computation of Gelman and Rubin's convergence diagnostic $\hat{R}$ (Gelman & Rubin, 1992). For each chain, we drew 100,000 samples from the posterior distribution, saving every 10th sample to remove autocorrelations in the Markov chain. If $\hat{R}$ was larger than 1.1, the simulation was excluded from the analysis.

Transparency and openness. All data and analysis code are available at https://osf.io/72uds. This study’s design and its analysis were not pre-registered.

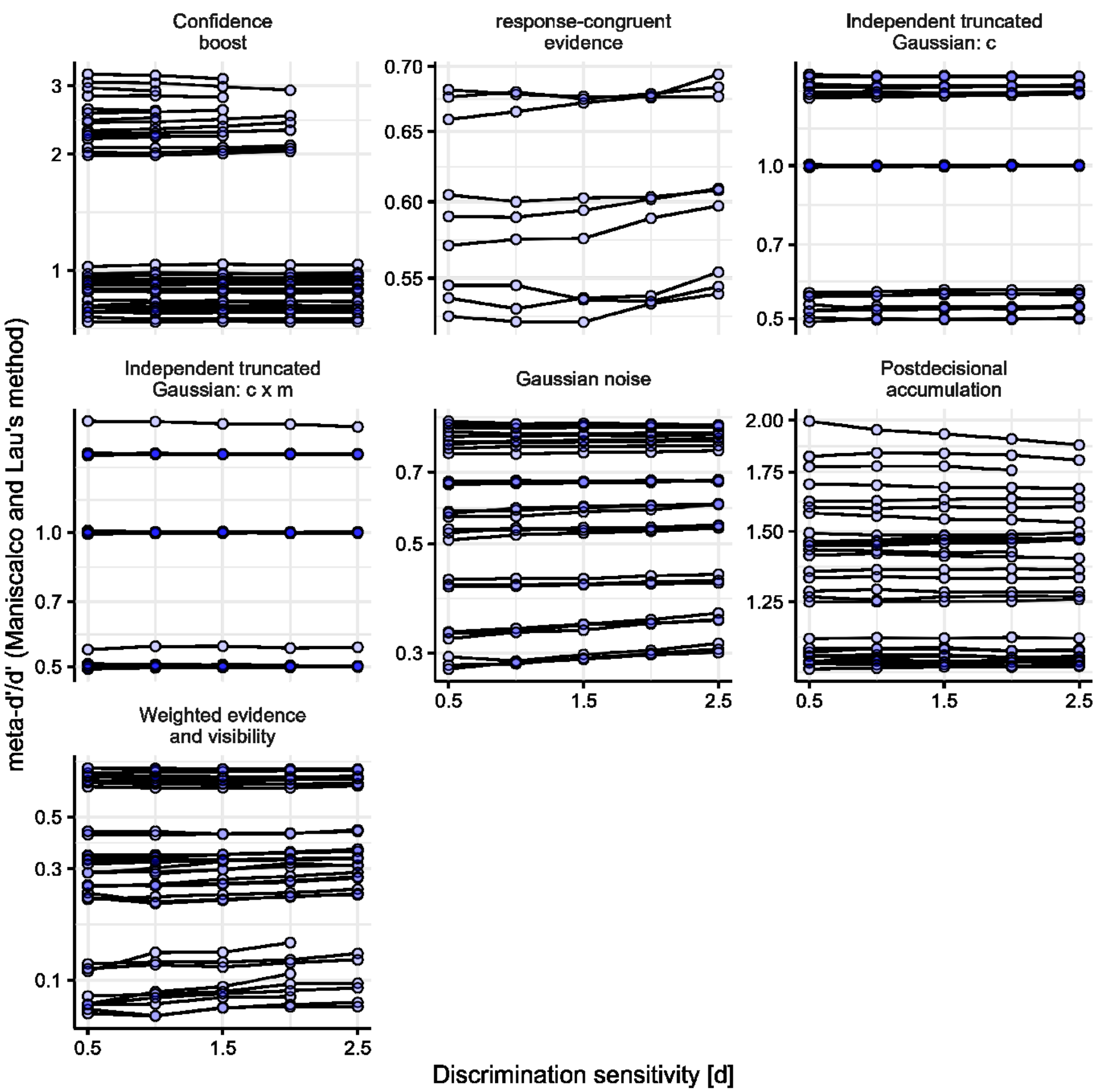
Results

Discrimination sensitivity

Fig. 4 shows the pattern of meta-d′/d′ as estimated using the conditioned maximum likelihood method proposed by Maniscalco and Lau (2012) as a function of the generative model underlying the simulated data and discrimination sensitivity. Meta-d′/d′ was not perfectly constant across different levels of discrimination sensitivity in any of the seven generative models. For the two independent truncated Gaussian models, meta-d′/d′ was associated with discrimination sensitivity only for a relatively small subset of simulations. In contrast, for the postdecisional accumulation model, the Gaussian noise model, the response-congruent evidence model, and the weighted evidence and visibility model, Fig. 4 shows multiple lines that have a non-zero slope, meaning that meta-d′/d′ depended on discrimination sensitivity for the majority of parameter sets.

Figure 4

Meta-d'-d' based on conditioned maximum likelihood estimation and model specification by Maniscalco and Lau, as function of discrimination sensitivity and generative model of confidence

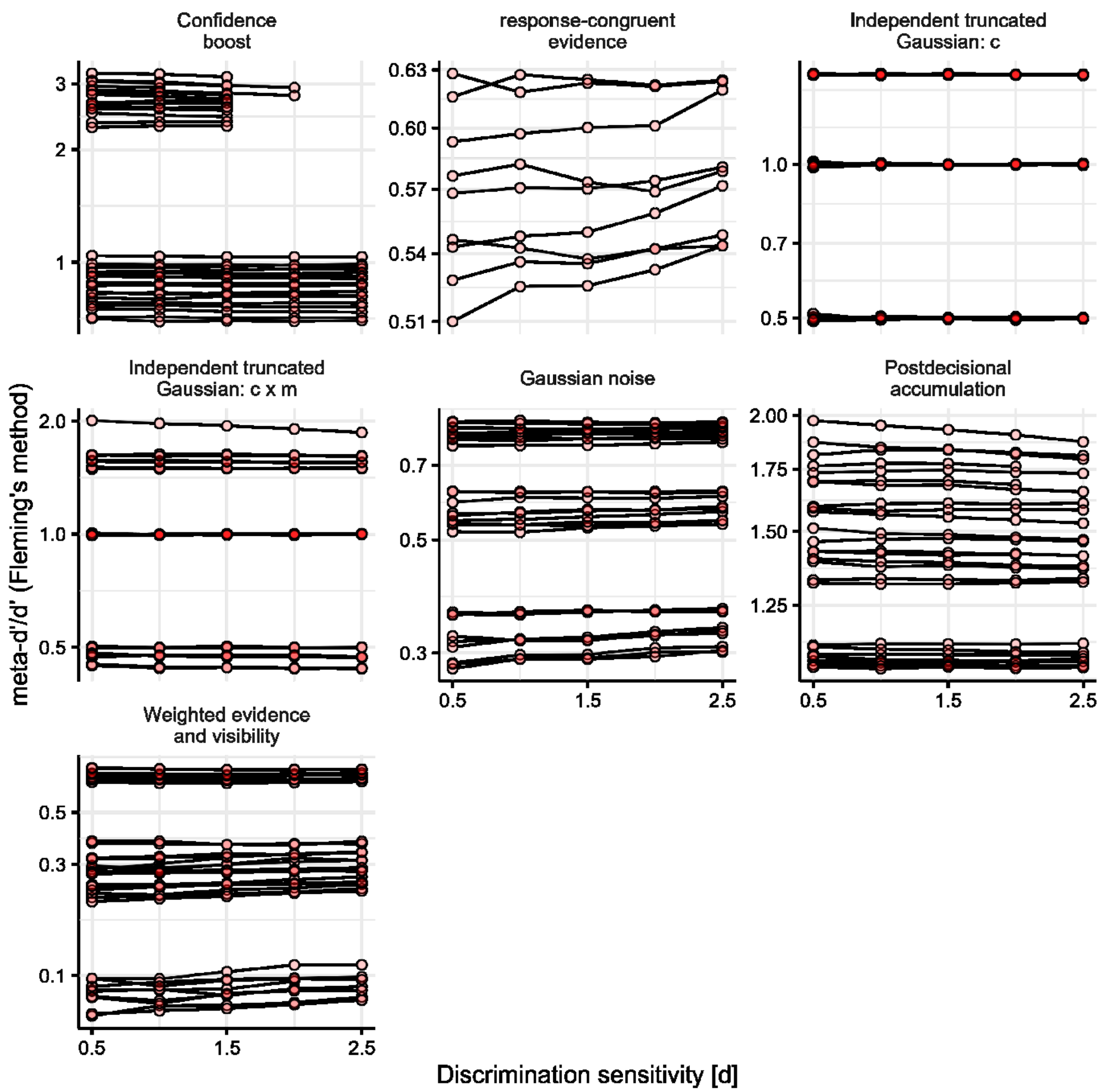


Note. Each dot represents one simulation with one combination of parameters. Lines connect simulations that differ only with respect to the parameter quantifying discrimination sensitivity and identical parameter sets otherwise. Lines parallel to the horizontal indicate that meta-d′/d′ is independent from discrimination sensitivity. Note that the y-Axes are different for each generative model of confidence.

Fig. 5 shows the pattern of meta-d′/d′ estimated using Fleming's Bayesian MCMC method, again as a function of the generative model underlying the simulated data and discrimination sensitivity. Meta-d′/d′ was constant across levels of discrimination performance when the data was generated according to the independent truncated Gaussian model with distributions truncated at the discrimination criterion c. When the same model was used but with distributions truncated at $c \times m$, there were some parameter sets where discrimination sensitivity affected meta-d′/d′. Again, for the postdecisional accumulation model, the Gaussian noise model, the response-congruent evidence model, and the weighted evidence and visibility model, discrimination sensitivity affected meta-d′/d′ ratios for a large number of parameter sets. When we repeated these analyses using conditioned maximum likelihood estimation but calculating the probability of confidence given stimulus and response following Fleming (2017), the results were visually indistinguishable from Fig. 5.

Figure 5

Meta-d'/d' based on Bayesian MCMC estimation and Fleming's model specification, as function of discrimination sensitivity and generative model of confidence



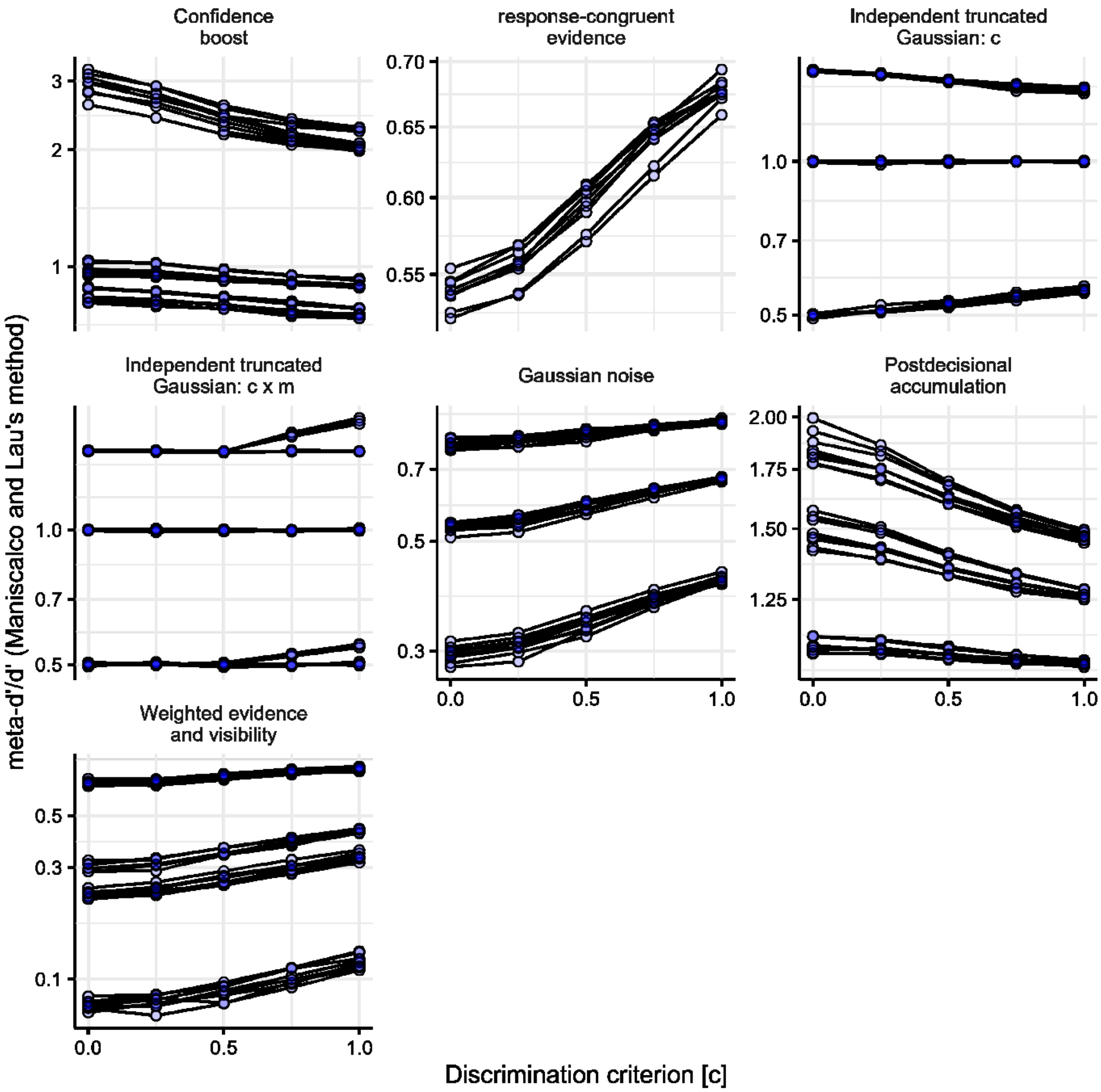
Note. Each dot represents one simulation. Lines connect simulations that differ only with respect to the parameter quantifying discrimination sensitivity and identical parameter sets otherwise. Lines parallel to the horizontal indicate that meta-d′/d′ is independent from discrimination sensitivity. Note that the y-Axes are different for each generative model of confidence.

Discrimination bias

The relationship between meta-d′/d′ and discrimination bias across different generative models is depicted in Fig. 6 for Maniscalco and Lau’s original conditioned maximum likelihood method and in Fig. 7 for Fleming's Bayesian MCMC method. Fig. 6 shows that meta-d′/d′ estimated using the original method depends on discrimination bias for each single generative model of confidence. Fig. 7 shows that meta-d′/d′ estimated using the Bayesian MCMC method is independent from discrimination bias only if the data is generated according to the independent truncated Gaussian model with the distributions truncated at the discrimination criterion. Again, meta-d′/d′ depends on the discrimination criterion according to all other generative models of confidence. Finally, when meta-d′/d′ was estimated using conditioned maximum likelihood estimation but using the model specification Fleming (2017), the results were the same as in Fig. 6.

Figure 6

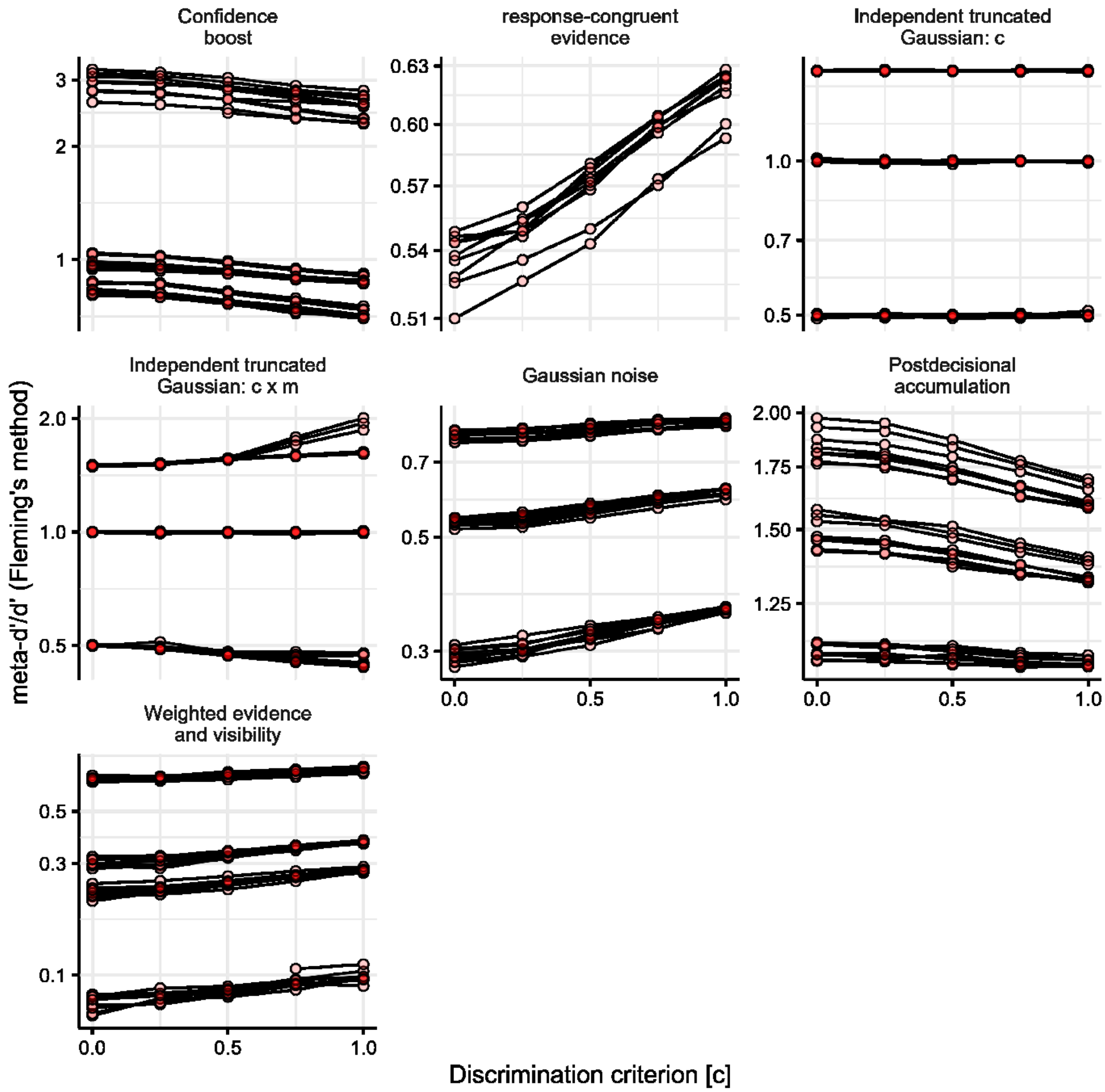
Meta-d'/d' based on conditioned maximum likelihood estimation and Maniscalco and Lau's model specification as function of discrimination bias and generative model of confidence



Note. Each dot represents one simulation. Lines connect simulations that differ only with respect to the parameter quantifying discrimination bias and identical parameter sets otherwise. Lines parallel to the horizontal indicate that meta-d′/d′ is independent from discrimination bias. Note that the y-Axes are different for each generative model of confidence.

Figure 7

Meta-d'/d' based on MCMC estimation and Fleming's model specification as function of discrimination bias and generative model of confidence



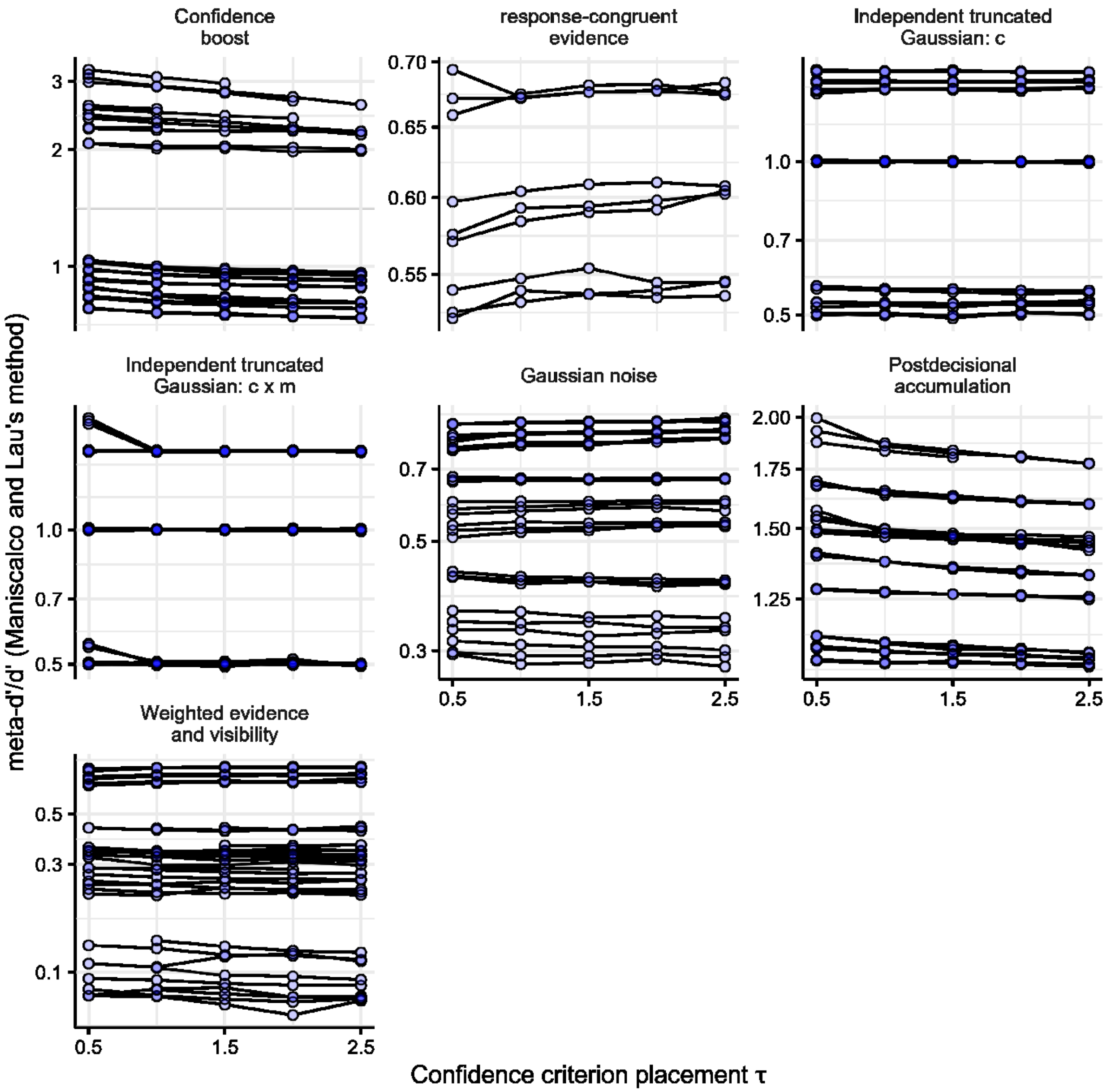
Note. Each dot represents one simulation. Lines connect simulations that differ only with respect to the parameter quantifying discrimination bias and identical parameter sets otherwise. Lines parallel to the horizontal indicate that meta-d′/d′ is independent from discrimination bias. Note that the y-axes are different for each generative model of confidence.

Confidence criteria

The relationship between meta-d′/d′ and confidence criterion placement across different generative models of confidence is depicted in Fig. 8 for Maniscalco and Lau’s original conditioned maximum likelihood method and in Fig. 9 for Fleming's Bayesian MCMC method. Fig. 8 shows that meta-d′/d′ estimated using the original method is never completely independent from confidence criterion placement. Nevertheless, for the two independent truncated Gaussian models, meta-d′/d′ was associated with confidence criterion placement for a relatively small subset of simulated parameter sets. Fig. 9 shows that meta-d′/d′ estimated using Fleming's method is independent from confidence criterion placement only if the data is generated according to the independent truncated Gaussian model with the distributions truncated at the discrimination criterion. For all other generative models of confidence, meta-d′/d′ depends on confidence criterion placement. Finally, when meta-d′/d′ was estimated using conditioned maximum likelihood estimation but with Fleming's model specification, the results were the same as in Fig. 9.

Figure 8

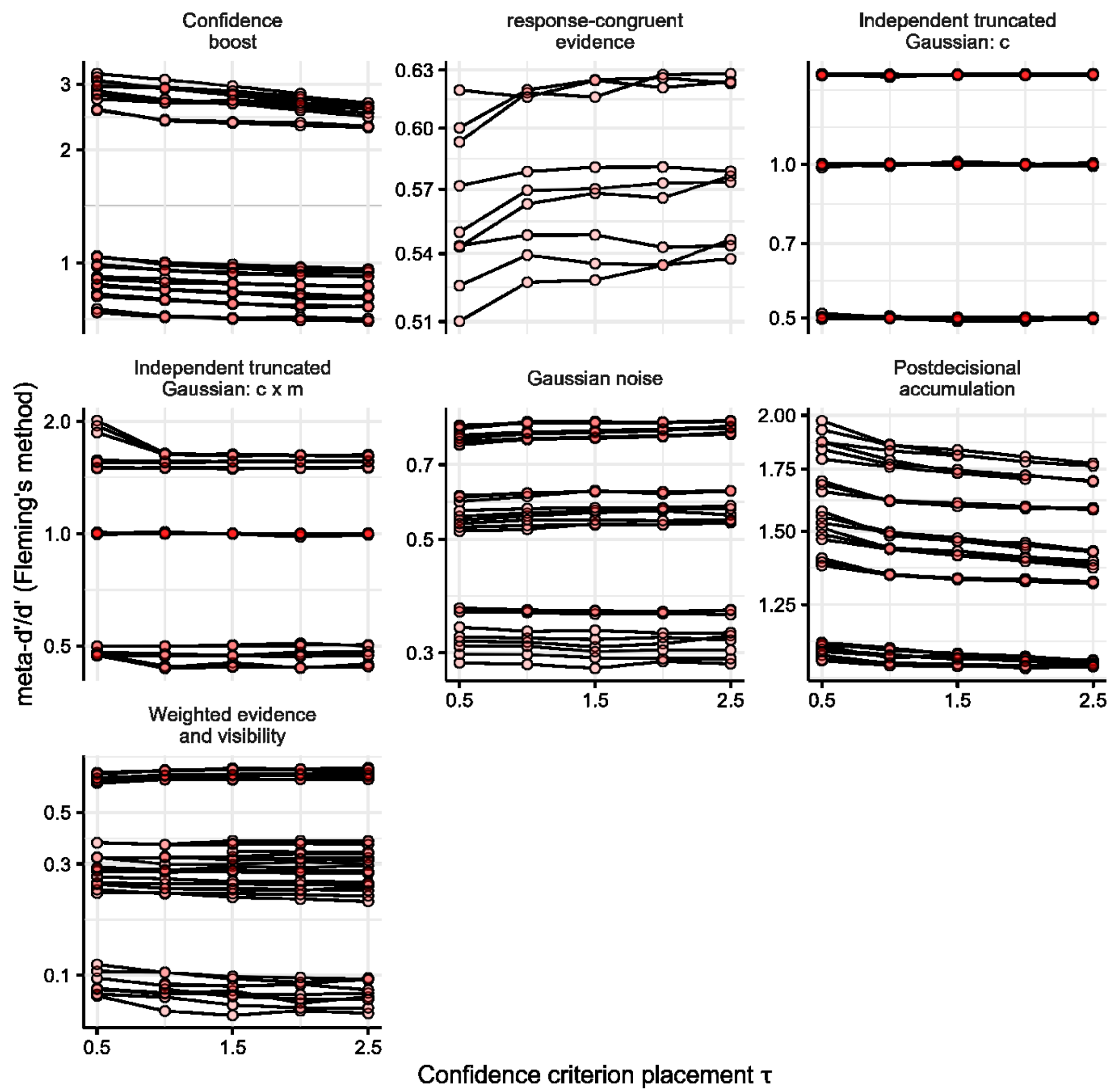
Meta-d'/d' based conditioned maximum likelihood estimation and Maniscalco and Lau's model specification as function of confidence criterion placement and generative model of confidence



Note. Each dot represents one simulation. Lines connect simulations that differ only with respect to the parameter determining confidence criterion placement and identical parameter sets otherwise. Lines parallel to the horizontal indicate that meta-d′/d′ is independent from confidence criterion placement. Note that the y-Axes are different for each generative model of confidence.

Figure 9

Meta-d'/d' based on MCMC estimation and Fleming's model specification as function of confidence criterion placement and generative model of confidence



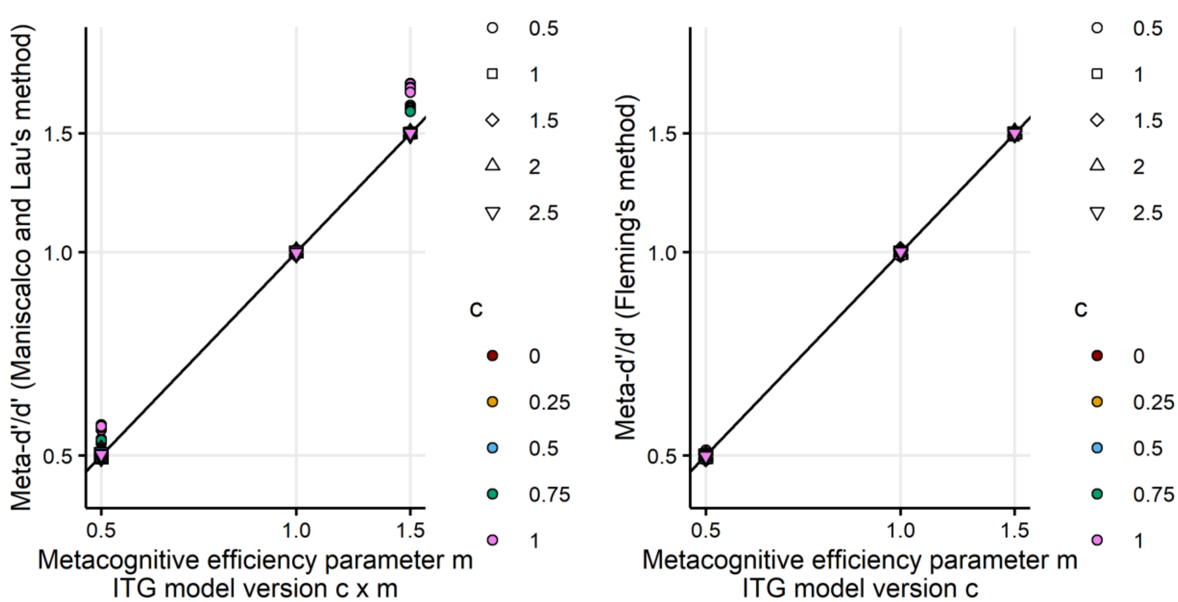
Note. Each dot represents one simulation. Lines connect simulations that differ only with respect to the parameter determining confidence criterion placement and identical parameter sets otherwise. Lines parallel to the horizontal indicate that meta-d′/d′ is independent from confidence criterion placement. Note that the y-Axes are different for each generative model of confidence.

Recovering metacognitive efficiency parameters

Finally, we investigated if estimates of meta-d′/d′ recover the metacognitive efficiency parameter m of the independent truncated Gaussian model. Specifically, meta-d′/d′ estimated using the original SDT model specification by Maniscalco and Lau (2014) was expected to recover the m parameter in the ITG model with the distribution truncated at the objective discrimination criterion c multiplied with m. Meta-d′/d′ estimated using the model specification by Fleming (2017) should recover the m parameter in the ITG model with the distribution truncated at c. Fig. 10 shows that meta-d′/d′ based on Bayesian MCMC estimation and Fleming's model specification indeed recovered the m parameter of the corresponding version of the ITG model. However, meta-d′/d′ using the model specification by Maniscalco and Lau (2014) did not always recover m in the corresponding ITG model. Specifically, meta-d′/d′ overestimated m when the discrimination criterion was at least .75 (i.e., a considerable bias for one of the two stimuli), when τ was 0.5 (i.e. liberal confidence criterion placement), and when m was either 0.5 or 1.5 (and thus metacognitive ability and perceptual ability were not the same).

Figure 10

Meta-d'/d' as a function of the metacognitive efficiency parameter m, discrimination bias parameter θ, and confidence criterion placement parameter τ.



Note. Left panel: ITG model with distributions truncated at the discrimination criterion c multiplied with m. Accordingly, meta-d′/d′ values on the y-axis were computed using the original method by Maniscalco and Lau (2014). Right panel: ITG model with distributions truncated at the discrimination criterion c. Accordingly, meta-d′/d′ values on the y-axis were computed using the Bayesian MCMC method by Fleming (2017). Colours indicate different objective discrimination criteria. Symbols indicate different placement of confidence criteria.

Discussion

Simulation 1 showed that meta-d′/d′ provides imperfect control over discrimination performance, discrimination bias, and confidence criteria: Only when the data were simulated according to the independent truncated Gaussian model with the distributions truncated at the discrimination bias, and when meta-d′/d′ was estimated using the model specification used by Fleming (2017), meta-d′/d′ was constant across discrimination performance, discrimination bias, and confidence criteria in all simulations. Notably, the control of discrimination sensitivity, bias, and confidence criteria is sensitive to the finer details of model specification: When we simulated data with distributions truncated at the discrimination bias multiplied by the metacognitive efficiency parameter, the generative model consistent with Maniscalco and Lau’s method, meta-d′/d′ based on Fleming’s model specification was no longer constant as a function of discrimination performance, discrimination bias, and confidence criteria across all simulated parameter sets. When the data were simulated according to one of the other generative models of confidence, meta-d′/d′ was associated with discrimination bias, discrimination sensitivity and confidence criterion placement for numerous simulations.

While Simulation 1 shows that meta-d′/d′ depends in principle on discrimination performance, discrimination bias and confidence criteria according to various different models of confidence, it is still unclear whether the effect is large enough to be relevant in practice. In particular, the contamination of meta-d′/d′ by discrimination sensitivity seemed to be relatively small compared to the contamination by discrimination bias and confidence criteria. However, in order to simulate the expected correlations between model parameters and meta-d′/d′ according to different confidence models, it is necessary to specify the distributions of the model parameters across subjects. Unfortunately, the sample sizes of previous modelling studies have been generally too small sample to reasonably estimate the distribution of model parameters across subjects.

Simulation 2

To investigate how the relationships observed in Simulation 1 may translate into plausible effect sizes, we fitted all seven models of confidence used in Simulation 1 to the data from Experiment 2 by Rouault et al. (2018), an open data set available from the confidence database (Rahnev et al., 2020). Rouault et al. (2018)'s data were chosen because a large sample is necessary for stable estimates of correlation coefficients (Schönbrodt & Perugini, 2013). We then used the parameter sets obtained by model fitting to simulate new data to estimate the correlation between meta-d′/d′ and discrimination sensitivity, discrimination bias and confidence criteria implied by each generative model of confidence.

Method

Experimental task

Rouault et al.'s data consists 497 subjects who participated in an online dot numerosity discrimination task with 210 trials per subject. In each trial, participants were presented with a fixation cross for 1 s. Two black boxes filled with differing numbers of randomly positioned white dots were then presented for 0.3 s. One box was always half-filled (313 dots out of 625 positions), while the other box contained an increased number of dots compared to the first box. The position of the box with the higher number of dots was pseudo-randomised across all trials. To maintain a constant level of performance during the experiment and across participants, a staircase was used to adapt the number of extra dots in the target box. The staircase started with a number of 70 extra dots and was a two-down one-up staircase procedure with equal step-sizes for steps up and down. The step-size was calculated in log-space, changing by ± 0.4 for the first 5 trials, ± 0.2 for the next 5 trials and ± 0.1 for the rest of the task. After 0.3 s, the dots disappeared, leaving the black boxes on screen until participants indicated which box had the higher number of dots by keyboard button press. Then, subjects were asked to report their confidence in their response on a 6-point rating scale with verbal descriptions (*certainly wrong*, *probably wrong*, *maybe wrong, maybe correct, probably correct, certainly correct*). A detailed description of the study is provided by Rouault et al. (2018).

Model fitting

All seven generative models of confidence used in Simulation 1 were fitted to the combined distributions of responses and confidence judgments separately for each single participant. The fitting procedure involved the following computational steps: First, the frequency of each confidence level was calculated for each of the two stimulus options and each of the response option. For each model, the set of parameters was determined that minimized the negative log-likelihood of the data given the model. For this purpose, we used a coarse grid search to identify five promising sets of starting values for the optimization procedure. Then, minimization of the negative log-likelihood was performed using a general SIMPLEX minimization routine (Nelder & Mead, 1965) for each set of starting values. To avoid local minima, the optimization procedure was restarted four times.

To assess the relative quality of the candidate models, we calculated the Bayes information criterion (Schwarz, 1978) and the AICc (Burnham & Anderson, 2002), a variant of the Akaike information criterion (Akaike, 1974) using the negative likelihood of each model fit with respect to each single participant and the trial number. For statistical testing, we compared the mean AICc and BIC using standard t-tests with *p*-values adjusted for multiple comparisons using Holm's correction.

Simulation

We simulated one new data set for each of the seven generative models of confidence, using the parameter sets we obtained during model fitting, using the same number of subjects as as in the empirical data and 10.000 trials per subject. Then, we estimated meta-d′/d′ two times for each simulated subject using conditioned maximum likelihood estimation, one time with Maniscalco and Lau's model specification, and one time with Fleming's model specification. Because meta-d′/d′ is not normally distributed (Rausch & Zehetleitner, 2023), we assessed the correlation between each parameter of each generative model and the logarithm of meta-d′/d′. We repeated the analysis using unstandardized linear regression slopes with centred parameters as predictors and log(meta-d′/d′) as criterion. All *p*-values were corrected for multiple comparisons using Holm's correction.

Results

Formal model comparisons

Formal model comparisons revealed that the best fits to the data were obtained by the two versions of the independent truncated Gaussian model, both in terms of AIC_c, and BIC. The difference between the two versions of the independent truncated Gaussian model was negligible, $M_{\Delta AIC} = M_{\Delta BIC} = 0.02$, $t(496) = 1.46$, $p = .290$. The fit of both independent truncated

Gaussian models was each significantly better than those of the five alternative models in terms of AIC and BIC, all *p's* < .001, although the mean difference was quite small, $M_{\Delta AIC}$'s and $M_{\Delta BIC}$'s ≥ -0.59, all *p's* < .001.

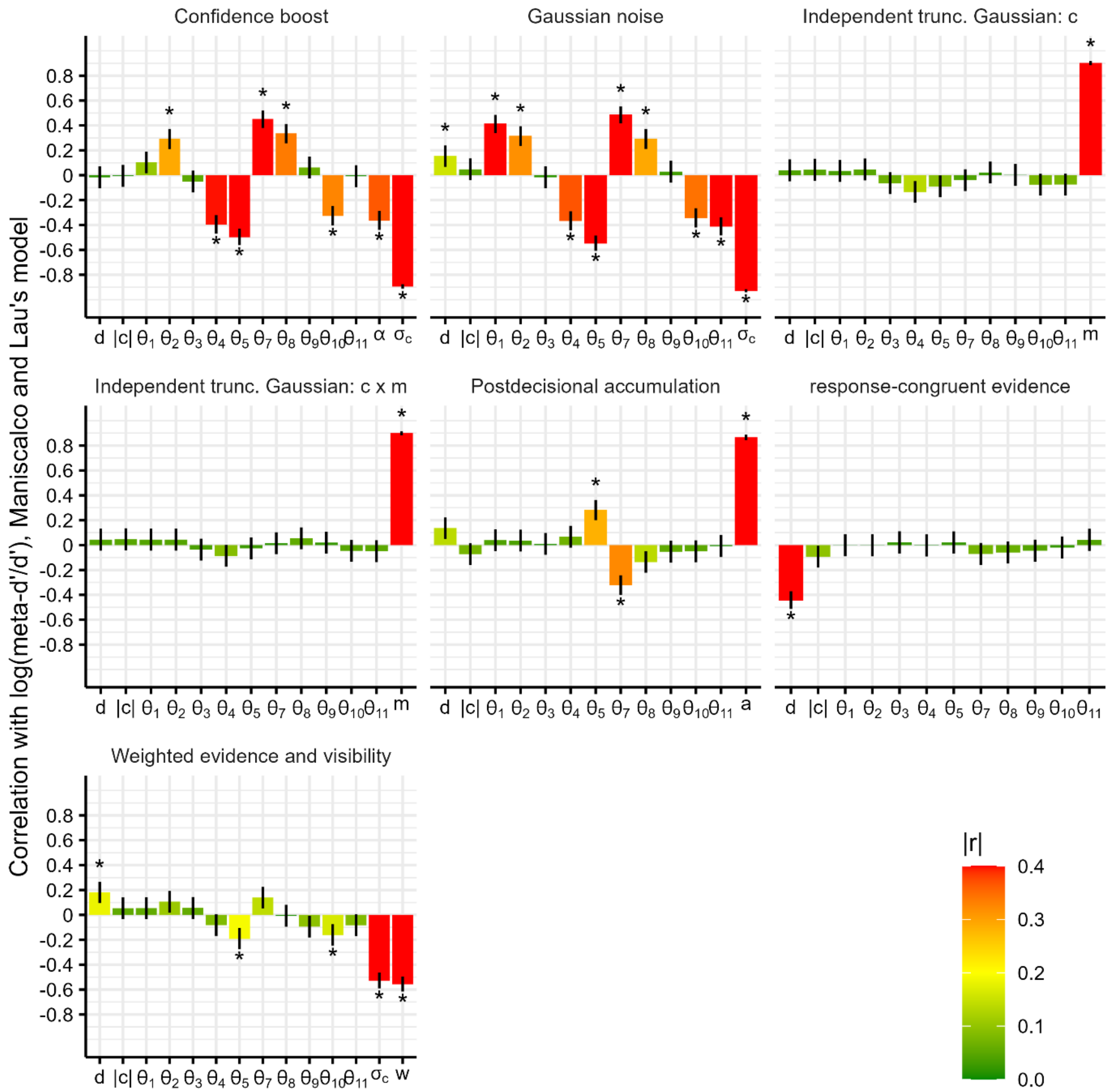
Correlations between model parameters and simulated meta-d′/d′

Supplementary Table 1 provides the correlation coefficients between each estimated parameter of the different confidence model and log(meta-d′/d′). Figs. 11 and 12 show that as expected, log(meta-d′/d′) is strongly correlated with all model parameters intended to reflect metacognitive efficiency, i.e. σ_c, m, a, and α. For the two versions of the independent gaussian truncated model, no significant correlation between log(meta-d′/d′) and discrimination sensitivity d, discrimination criterion c, or any of the ten confidence criteria was observed, independently from the specification of the hypothetical SDT model underlying meta-d′/d′. However, we found a significant large correlation between discrimination sensitivity d and log(meta-d′/d′) for the response-congruent evidence model and a medium-sized correlation between discrimination sensitivity d and log(meta-d′/d′) for the weighted evidence and visibility model. For the Gaussian noise model, a moderate correlation between d and log(meta-d′/d′) was significant only when meta-d′/d′ was estimated using Maniscalco and Lau's model specification, but not for Fleming's model specification. On the contrary, for the postdecisional accumulation model, the correlation between d and log(meta-d′/d′) was significant only when meta-d′/d′ was estimated based on Fleming's model, but not with Maniscalco and Lau's model. A significant medium-sized correlation between log(meta-d′/d′) and discrimination bias c was detected only for the response-congruent evidence model when meta-d′/d′ was estimated using Fleming's model specification. Concerning confidence criteria, we found a very strong correlation between log(meta-d′/d′) and six out of ten confidence criteria for the confidence boost model, seven out of ten for the Gaussian noise model, and two out of ten for the postdecisional accumulation model. In addition, we detected medium-sized correlations between two confidence criteria in the weighted evidence and visibility model.

The analysis of regression slopes revealed that for the confidence boost model and the Gaussian noise model, there were only small changes in meta-d′/d′ as a function of confidence criteria, but these changes were very consistent across subjects, resulting in many significant small effects. For the other models and parameters, the interpretation was essentially the same as in the correlation analysis (see Supplementary Table 2).

Figure 11

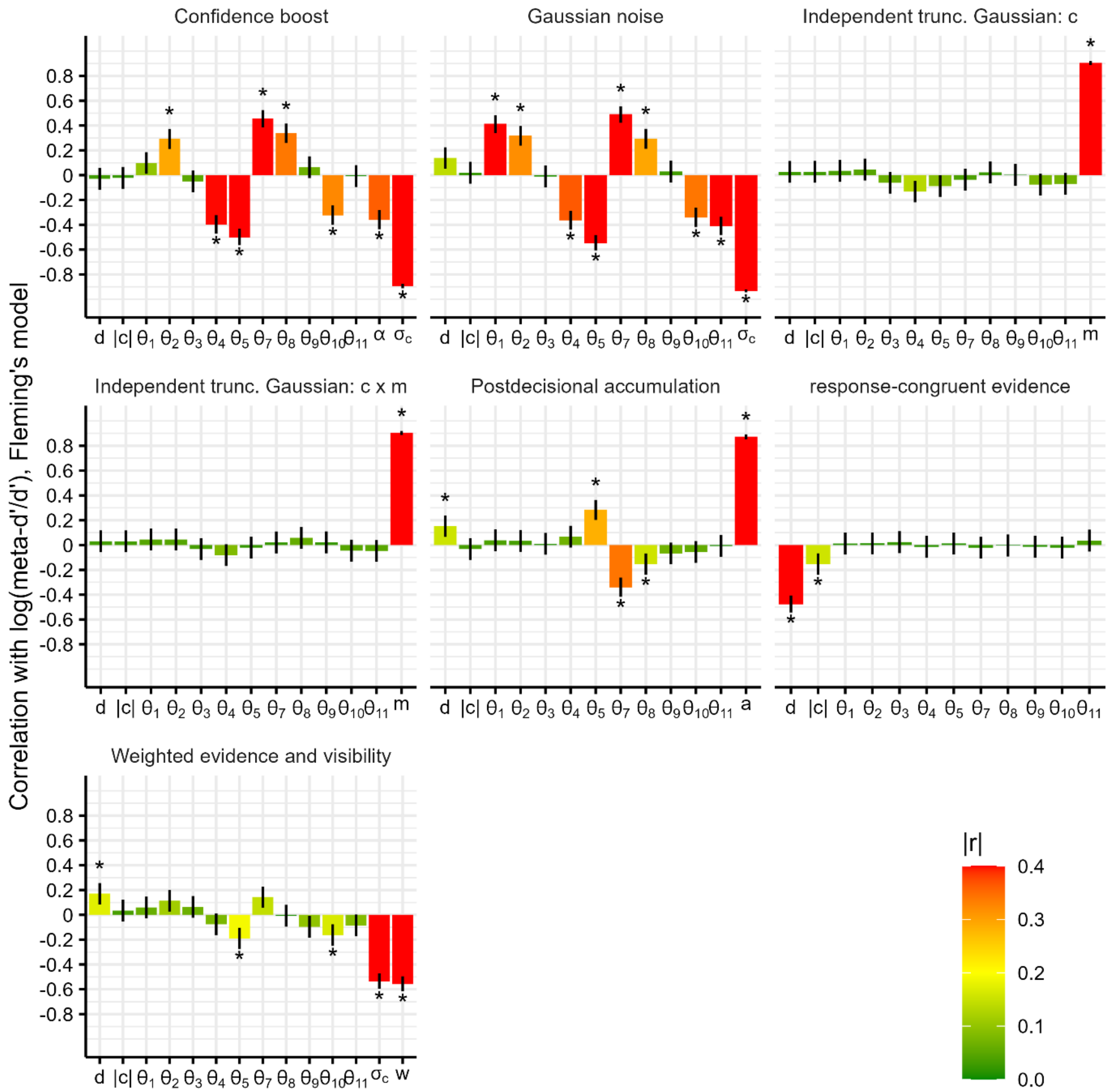
Correlation between meta-d'/d' estimated using Maniscalco and Lau's model specification and model parameters estimated from Rouault et al. (2018)'s Exp. 2 as a function of different generative models of confidence.



Note. Error bars indicate 95% CI.

Figure 12

Correlation between log-transformed meta-d'/d' estimated using Fleming's model specification and model parameters model parameters estimated from Rouault et al. (2018)'s Exp. 2 as a function of different generative models of confidence.



Note. Error bars indicate 95% CI.

Discussion

Fitting different models of confidence to Rouault et al. (2018)'s data showed that the two versions of the independent truncated Gaussian model provide a reasonable fit to confidence in a dot numerosity discrimination task. Importantly, the model comparisons reported in the present study should be only interpreted as preliminary, because the data set only included 200 trials per subject, which is much smaller than the norm in modelling studies. It should also be noted that the statistical properties of different experimental tasks may be quite different, suggesting that the observation that ITG performs well in one data set does not imply that ITG will also perform well in other experimental tasks. Nevertheless, we think that ITG should be considered as a series candidate model in future studies and should be routinely included in future comparisons of confidence models.

The simulation using the parameters of the independent truncated Gaussian model obtained during model fitting showed that both versions of meta-d′/d′ were independent of discrimination sensitivity, discrimination bias, and confidence criteria, suggesting that the differences between the two versions of the independent truncated Gaussian model are small enough not to be practically relevant, at least with distributions of parameters as observed in this particular experiment. However, for each of the five alternative models of confidence, we found at least one strong correlation with either discrimination sensitivity or one of the confidence criteria. The correlations with discrimination sensitivity parameters are noteworthy because Rouault et al. used a staircase to keep accuracy constant. This means that staircases still leave enough variance in discrimination sensitivity parameters to produce a large correlation with discrimination sensitivity for the response-congruent evidence model, medium-sized correlations for the weighted evidence and visibility model, or small-to-medium correlations for the gaussian noise model and the postdecisional accumulation model.

General discussion

The results of the present study suggests that whether or not meta-d′/d′ provides control over discrimination performance, discrimination bias, and confidence criteria strongly depends on the generative model of confidence: Only when the data was simulated according to the independent truncated Gaussian model (ITG) with the distributions truncated at the discrimination bias, and when meta-d′/d′ was estimated using the model specification used by Fleming (2017), meta-d′/d′ was perfectly constant across discrimination performance, discrimination bias, and confidence criteria across all simulations. When we simulated data using the parameters estimated from Rouault et al. (2018)'s Exp. 2, no difference between the two versions of ITG were observed, suggesting that the difference between the two versions of ITG may not always be relevant in practice. However, when the data was simulated not using ITG, but the Gaussian noise model, the postdecisional accumulation model, the weighted evidence and visibility model, the confidence boost model, or the response-congruent evidence model, meta-d′/d′ depended on discrimination sensitivity, discrimination bias, and confidence criterion placement for many simulations. Simulations using parameters obtained by fitting empirical data showed that the expected correlations between meta-d′/d′ and model parameters vary widely across different generative model of confidence and specific parameters. Nevertheless, for each generative model other than ITG, there was at least one medium-sized correlation with either discrimination sensitivity or one of the confidence criteria, suggesting that meta-d′/d′ is associated with discrimination sensitivity and confidence criteria under realistic assumptions about model parameters.

Relation between meta-d′/d′ and generative models of confidence

Meta-d′/d′ has been considered to rely only on the assumption of a specific cognitive architecture underlying the discrimination decision, but to be free from assumptions about the decision variable underlying the confidence decision (Maniscalco & Lau, 2014). In contrast, the main finding of the present study is that meta-d′/d′ is in fact not free from assumptions about the generative model underlying confidence judgments. The reason is that meta-d′/d′ depends on discrimination sensitivity, discrimination bias, and confidence criteria when the data is simulated according to the Gaussian noise model, the weighted evidence and visibility model, the confidence boost model, the postdecisional accumulation model or the response-congruent evidence model. Previous studies revealed two additional models where meta-d′/d′ is confounded, the Bayesian beta-distributed noise model (Guggenmos, 2021) and the lognormal noise model (Shekhar & Rahnev, 2021). Importantly, the present study exceeds those studies in showing that generative models where meta-d′/d′ is contaminated by discrimination sensitivity, discrimination bias and confidence criteria not only exist, but the same result is obtained according to most generative models of confidence. Meta-d′/d′ succeeds in controlling for discrimination sensitivity, discrimination bias and confidence criteria when the data is generated according to the independent truncated gaussian model. Thus, it seems that the control meta-d′/d′ provides is highly specific to the independent truncated Gaussian model. Our findings are consistent with the assertion that discrimination sensitivity, discrimination bias and confidence criteria can only be controlled based on estimating the underlying generative model of confidence (Guggenmos, 2022). We cannot prove that no generative model other than ITG exists where meta-d′/d′ performs satisfactorily. However, the control over discrimination sensitivity, discrimination bias, and confidence criteria fails for a large variety of different generative models, which is why it is reasonable to assume that meta-d′/d′ is unlikely to provide effective control in other models which were not examined so far. Overall, this means that meta-d′/d′ from now on should be regarded as a model-based measure of metacognitive efficiency, and researchers who consider using meta-d′/d′ need to ascertain if their data can be adequately described by ITG.

Evidence for the independent truncated Gaussian model?

Because the adequacy of meta-d′/d′ depends on the assumption of ITG as generative model, the question is raised if ITG is a decent models of human confidence judgments. Our analysis of the data of Rouault et al. (2018) is to our knowledge the first (albeit preliminary) evidence that data sets exists which are adequately described by ITG. Unfortunately, previous studies comparing generative models of confidence did not make the link between meta-d′/d′ and generative model of confidence, which is why ITG has not been included into formal model comparisons previously (e.g. Maniscalco & Lau, 2016; Rausch et al., 2018, 2020, 2021; Shekhar & Rahnev, 2021, 2022). Future modelling studies are necessary to investigate how frequently ITG is an adequate description of human confidence. However, there is more evidence for some qualitative predictions of ITG. According to ITG, confidence judgments are subject to a response-congruent confirmation bias because it is impossible to sample a confidence decision variable that contradicts the discrimination decision. In accordance with ITG, previous studies reported that observers' tend to neglect contradictory evidence when they report confidence (Peters et al., 2017; Samaha et al., 2016; Zylberberg et al., 2012), although no evidence for a response-congruent confirmation bias was observed in other experimental paradigms (Rausch et al., 2020; Shekhar & Rahnev, 2022), suggesting a response-congruent confirmation bias many not be a universal feature of human confidence across paradigms. However, there are multiple mathematical ways to represent bias in favour of response-congruent evidence. When we implemented a response-congruent evidence bias in a different way, resulting in the model we refer to as response-congruent evidence model, meta-d′/d′ very strongly correlated with discrimination sensitivity. This finding implies that it is not sufficient that the generative process underlying the confidence data is characterised by a similar conceptual idea as ITG - if meta-d′/d′

is to control for by discrimination sensitivity, discrimination bias, and confidence criteria, ITG must be (at least a close approximation of) the generative model of the data.

An important limitation shared between ITG and all alternative models investigated in the present study is that the dynamics of the decision process is not accounted for. This is problematic because there is a large body of evidence that confidence judgments depend on the dynamics of decision making (Pleskac & Busemeyer, 2010). Specifically, Pleskac and Busemeyer (2010) showed that when participants are under time pressure when making the decision, metacognitive efficiency is increased. Moran et al. (2015) showed that confidence judgments are related to the reaction time of confidence judgments. Last but not least, there is on average medium-sized correlation between confidence judgments and reaction time across a wide range of studies (Rahnev et al., 2020). Given the close relationship between decision dynamics and confidence, it may be more apt to model confidence with sequential sampling models rather than signal detection theory (Desender et al., 2022; Hellmann et al., 2023; Pereira et al., 2021; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009, 2013; Reynolds et al., 2020).

Measuring metacognitive efficiency using meta-d′/d′

The findings of the present study imply that whenever the independent truncated Gaussian model is a good description of the data, meta-d′/d′ will be the appropriate measure of metacognitive efficiency. However, without any information about the generative model underlying confidence judgments, researchers should not assume that by using meta-d′/d′ to measure metacognitive efficiency, a potential contamination by discrimination sensitivity, discrimination bias, or confidence criteria has been ruled out. We recommend to use meta-d′/d′ only for tasks where the independent truncated Gaussian model is a suitable description of the data. There is a limited set of experimental tools available to reduce the potential impact of discrimination sensitivity, discrimination bias, and confidence criteria when measuring metacognitive efficiency using meta-d′/d′. To control for discrimination sensitivity, researchers have used staircases to keep task performance within a specific range (Rahnev & Fleming, 2019). However, Simulation 2 suggests that staircases are not sufficient to control for discrimination sensitivity if the data is generated according to the weighted evidence and visibility model or the response-congruent evidence model. It might be possible to reduce the impact of discrimination criteria and confidence criteria by careful instructions and training with the task, although it is unlikely that instruction and training is sufficient to eliminate the effect of criteria.

Measuring metacognitive efficiency by meta-d′/d′ is also problematic because meta-d′/d′ does not take the dynamics of the decision process into account. Consequently, properties of the dynamical decision process such as response caution might be misinterpreted as effects on metacognitive efficiency (Desender et al., 2022). Overall, the findings of the present study combined with other recent studies (Desender et al., 2022; Guggenmos, 2021; Shekhar & Rahnev, 2021) imply that without any additional information, meta-d′/d′ cannot be unambiguously interpreted in terms of metacognitive efficiency, suggesting that a reanalysis of previously published studies using meta-d′/d′ and possibly a critical reinterpretation is necessary.

Alternatives to meta-d′/d′ for measuring metacognitive efficiency

Whenever ITG is not a decent description of confidence in a particular study, researchers need an alternative to meta-d′-d′ to measure metacognitive efficiency. Traditionally, metacognition has been assessed using measures that also do not explicitly rely on specific generative models of confidence, such as gamma correlation coefficients (Nelson, 1984), confidence slopes (Yates, 1990), phi correlations (Rounis et al., 2010), or area under type 2 ROC curves (Fleming et al., 2010). However, none of these measures is designed to control for discrimination performance and thus, by definition, none of these measures are measures of metacognitive efficiency.

There are several model-based alternative measures of metacognitive efficiency: First, one available method is to fit a lognormal noise model, in which metacognitive ability is quantified by the lognormal noise parameter σ_{meta} (Shekhar & Rahnev, 2021, 2022). The lognormal noise model provides a decent account for confidence in a low contrast orientation discrimination task as well as a letter numerosity discrimination task (Shekhar & Rahnev, 2022). Second, in two-alternative forced choice confidence paradigms, it is possible to quantify metacognitive efficiency using the confidence boost model (Mamassian & de Gardelle, 2021). The measure of metacognitive efficiency η is computed by dividing the variance of the confidence noise of a hypothetical ideal observer by the variance of confidence noise estimated for the participant. Besides, two-alternative forced choice confidence paradigms may be an attractive way to eliminate the impact of confidence criteria (Barthelmé & Mamassian, 2009). Finally, relying on two-stage signal detection theory (Pleskac & Busemeyer, 2010; Yu et al., 2015), Desender et al. (2022) proposed the v-ratio to measure metacognitive efficiency. The v-ratio divides the drift rate estimated from confidence judgments by the drift rate estimated from discrimination responses and reaction time.

Notably, just as meta-d′/d′ is only a good measure of metacognitive efficiency when the data confirm to the independent truncated Gaussian model, σ_{meta}, η, and v-ratio are expected to control for discrimination sensitivity, discrimination bias and confidence criteria only when the data confirm to the corresponding generative model. To our knowledge, it has not yet been investigated how sensitive σ_{meta}, η, and v-ratio are to a contamination from discrimination sensitivity, discrimination bias and confidence criteria are when generative model underlying confidence judgment is varied. The findings of the present study are consistent with the view that measures of metacognitive efficiency provide control over discrimination sensitivity, discrimination bias and confidence criteria only if the generative model of confidence is correctly identified and the corresponding measure of metacognitive efficiency is used (Guggenmos, 2022). Unfortunately, for the time being, there is no consensus about the computational principles underlying confidence judgments (Rahnev et al., 2022). This means that a good practice for future studies will be to first use cognitive modelling to identify the generative model underlying confidence judgments in a specific paradigm empirically, and then use the corresponding model-based measure of metacognitive efficiency (Guggenmos, 2021; Mamassian & de Gardelle, 2021; Shekhar & Rahnev, 2021). When data in a specific task is well accounted for by the independent truncated Gaussian model, meta-d′/d′ is the appropriate way to measure metacognitive efficiency. However, when data is better described by an alternative model of confidence, researchers need to use a measure of metacognitive efficiency that corresponds to the model that is the best explanation of the data. Because researchers have implicitly fitted versions of the independent truncated Gaussian model all along when they used meta-d′/d′, it does not seem too far-fetched that researchers will begin to regularly fit alternative generative models of confidence as well. It will be necessary to develop open and easy-to-use software packages to make fitting a variety of confidence models available to a larger part of the field (e.g., Rausch & Hellmann, 2023). Sometimes it will be impossible to identify the true generative model underlying confidence judgments for a specific data set, either because the number of trials is too low or because of model mimicry. In these cases, it will be prudent to perform a robustness analysis to show that the results of the study do not depend on specific analysis decisions (Gelman & Loken, 2014; Steegen et al., 2016). This means that the modelling analysis needs to be repeated with all models of confidence that cannot be ruled out empirically to show that results are robust across models of confidence.

It is very difficult, and perhaps impossible, to come up with a novel measure of metacognitive efficiency with all the attractive properties that meta-d′/d′ was supposed to have, i.e., controlling for discrimination sensitivity, discrimination bias, and confidence criteria without requiring a specific generative model of confidence. The present study does not rule out the possibility that a future study will be able to find such a measure. However, given the results of the present study, we are sceptical that such a measure can ever be found; we recommend rigorous testing of whether any newly proposed measure of metacognitive efficiency effectively controls for discrimination performance, discrimination bias, and confidence criteria.

Conclusion

We showed that meta-d′/d′ is not free from assumptions about the generative model underlying confidence judgments. Only if the data is generated according to the independent truncated gaussian model, meta-d′/d′ guarantees control over discrimination performance, discrimination bias, and confidence criteria. The control fails according to a wide range of alternative generative models of confidence; the expected correlation with discrimination sensitivity and confidence criteria varies across alternative generative model but can be very large. Consequently, researchers who want to measure metacognitive efficiency using meta-d′/d′ need to examine if their data can be reasonably described by the independent truncated Gaussian model.

References

Adler, W. T., & Ma, W. J. (2018). Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLOS Computational Biology*, *14*(11), e1006572. https://doi.org/10.1371/journal.pcbi.1006572

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. https://doi.org/10.1371/journal.pcbi.1004519

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE transactions on automatic control*, *AC-19*(6), 716–723. https://doi.org/10.1007/978-1-4612-1694-0_16

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–552. https://doi.org/10.1037/a0033268

Barthelmé, S., & Mamassian, P. (2009). Evaluation of Objective Uncertainty in the Visual System. *PLoS Computational Biology*, *5*(9), e1000504. https://doi.org/10.1371/Citation

Bhome, R., McWilliams, A., Price, G., Poole, N. A., Howard, R. J., Fleming, S. M., & Huntley, J. D. (2022). Metacognition in functional cognitive disorder. *Brain Communications*, *4*(2), fcac041. https://doi.org/10.1093/braincomms/fcac041

Boldt, A., Blundell, C., & De Martino, B. (2019). Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of Consciousness*, *2019*(1), niz004. https://doi.org/10.1093/nc/niz004

Boundy-Singer, Z. M., Ziemba, C. M., & Goris, R. L. T. (2022). Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour*, *7*(1), 142–154. https://doi.org/10.1038/s41562-022-01464-x

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2. Aufl.). Springer.

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage*, *73*, 80–94. https://doi.org/10.1016/j.neuroimage.2013.01.054

Clarke, F. R., Birdsall, T. G., & Tanner, W. P. (1959). Two Types of ROC Curves and Definitions of Parameters. *The Journal of the Acoustical Society of America*, *31*(5), 629–630. https://doi.org/10.1121/1.1907764

Culot, C., Fantini-Hauwel, C., & Gevers, W. (2021). The influence of sad mood induction on task performance and metacognition. *Quarterly Journal of Experimental Psychology*, *74*(9), 1605–1614. https://doi.org/10.1177/17470218211004205

Desender, K., Donner, T. H., & Verguts, T. (2021). Dynamic expressions of confidence within an evidence accumulation framework. *Cognition*, *207*(104522), 1–11. https://doi.org/10.1016/j.cognition.2020.104522

Desender, K., Vermeylen, L., & Verguts, T. (2022). Dynamic influences on static measures of metacognition. *Nature Communications*, *13*(1), 1–30. https://doi.org/10.1038/s41467-022-31727-0

Drescher, L. H., Van den Bussche, E., & Desender, K. (2018). Absence without leave or leave without absence: Examining the interrelations among mind wandering, metacognition and cognitive control. *PLOS ONE*, *13*(2), e0191639. https://doi.org/10.1371/journal.pone.0191639

Fischer, H., & Said, N. (2021). Importance of domain-specific metacognition for explaining beliefs about politicized science: The case of climate change. *Cognition*, *208*, 104545. https://doi.org/10.1016/j.cognition.2020.104545

Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *1*, 1–14. https://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. https://doi.org/10.1037/rev0000045

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience*, *8*(443), 1–9. https://doi.org/10.3389/fnhum.2014.00443

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543. https://doi.org/10.1126/science.1191883

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.

Gelman, A., & Loken, E. (2014). The statistical Crisis in science. *American Scientist*, *102*(6), 460–465. https://doi.org/10.1511/2014.111.460

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–511.

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*, *7*(1), 1–14. https://doi.org/10.1093/nc/niab040

Guggenmos, M. (2022). Reverse engineering of metacognition. *eLife*, *11*, 1–29. https://doi.org/10.7554/eLife.75420

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), 5602. https://doi.org/10.1038/s41598-018-23936-9

Hellmann, S., Zehetleitner, M., & Rausch, M. (2023). Simultaneous modeling of choice, confidence, and response time in visual perception. *Psychological Review*. https://doi.org/10.1037/rev0000411

Kristensen, S. B., Sandberg, K., & Bibby, B. M. (2020). Regression methods for metacognitive sensitivity. *Journal of Mathematical Psychology*, *94*(102297), 1–17. https://doi.org/10.1016/j.jmp.2019.102297

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory. A user's guide*. Lawrence Erlbaum Associates.

Mamassian, P., & de Gardelle, V. (2021). Modeling Perceptual Confidence and the Confidence Forced-Choice Paradigm. *Psychological Review*, 1–23. https://doi.org/10.1037/rev0000312

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

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

Maniscalco, B., & Lau, H. C. (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 S. M. Fleming & C. D. Frith (Hrsg.), *The Cognitive Neuroscience of Metacognition* (S. 25–66). Springer. https://doi.org/10.1007/978-3-642-45190-4_3

Maniscalco, B., McCurdy, L. Y., Odegaard, B., & Lau, H. (2017). Limited Cognitive Resources Explain a Trade-Off between Perceptual and Metacognitive Vigilance. *The Journal of Neuroscience*, *37*(5), 1213–1224. https://doi.org/10.1523/JNEUROSCI.2271-13.2016

Maniscalco, B., Peters, M. A. K., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception & Psychophysics*, *78*, 923–937. https://doi.org/10.3758/s13414-016-1059-x

Mazancieux, A., Dinze, C., Souchay, C., & Moulin, C. J. A. (2020). Metacognitive domain specificity in feeling-of-knowing but not retrospective confidence. *Neuroscience of Consciousness*, *2020*(1), niaa001. https://doi.org/10.1093/nc/niaa001

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. https://doi.org/10.1016/j.cogpsych.2015.01.002

Muthesius, A., Grothey, F., Cunningham, C., Hölzer, S., Vogeley, K., & Schultz, J. (2022). Preserved metacognition despite impaired perception of intentionality cues in schizophrenia. *Schizophrenia Research: Cognition*, *27*, 100215. https://doi.org/10.1016/j.scog.2021.100215

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313. https://doi.org/10.1093/COMJNL/7.4.308

Nelson, T. O. (1984). A Comparison of Current Measures of the Accuracy of Feeling-of-Knowing Predictions. *Psychological Bulletin*, *95*(1), 109–133.

Odegaard, B., Chang, M. Y., Lau, H., & Cheung, S.-H. (2018). Inflation versus filling-in: Why we feel we see more than we actually do in peripheral vision. *Phil. Trans. R. Soc. B*, *373*, 20170345. https://doi.org/10.1098/rstb.2017.0345

Odegaard, B., Grimaldi, P., Cho, S. H., Peters, M. A. K., Lau, H., & Basso, M. A. (2018). Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proceedings of the National Academy of Sciences*, *115*(7), E1588–E1597. https://doi.org/10.1073/pnas.1711628115

Pereira, M., Megevand, P., Tan, M. X., Chang, W., Wang, S., Rezai, A., Seeck, M., Corniola, M., Momjian, S., Bernasconi, F., Blanke, O., & Faivre, N. (2021). Evidence accumulation relates to perceptual consciousness and monitoring. *Nature Communications*, *12*(1), 3261. https://doi.org/10.1038/s41467-021-23540-y

Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., Halgren, E., & Lau, H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, *1*(0139), 1–21. https://doi.org/10.1038/s41562-017-0139

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 171–212. https://doi.org/10.1109/TIT.1954.1057460

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. https://doi.org/10.1037/a0019737

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

Pollack, I. (1959). On Indices of Signal and Response Discriminability. *Journal of the Acoustical Society of America*, *31*, 1031.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rahnev, D., Balsdon, T., Charles, L., de Gardelle, V., Denison, R., Desender, K., Faivre, N., Filevich, E., Fleming, S. M., Jehee, J., Lau, H., Lee, A. L. F., Locke, S. M., Mamassian, P., Odegaard, B., Peters, M. A. K., Reyes, G., Rouault, M., Sackur, J., … Zylberberg, A. (2022). Consensus Goals in the Field of Visual Metacognition. *Perspectives on Psychological Science*, *17*(6), 1746–1765. https://doi.org/10.1177/174569162210756

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., Bang, J. W., Bègue, I., Birney, D. P., Brady, T. F., Calder-Travis, J., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T. C., … Zylberberg, A. (2020). The Confidence Database. *Nature Human Behaviour*, *4*, 317–325. https://doi.org/10.1038/s41562-019-0813-1

Rahnev, D., & Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness*, *5*(1), 1–9. https://doi.org/10.1093/nc/niz009

Ratcliff, R., & Starns, J. J. (2009). Modeling Confidence and Response Time in Recognition Memory. *Psychological Review*, *116*(1), 59–83. https://doi.org/10.1037/a0014086

Ratcliff, R., & Starns, J. J. (2013). Modeling Confidence Judgments, Response Times, and Multiple Choices in Decision Making: Recognition Memory and Motion Discrimination. *Psychological Review*, *120*(3), 697–719. https://doi.org/10.1037/a0033152

Rausch, M., & Hellmann, S. (2023). *statConfR: Models of Decision Confidence and Metacognition* (0.0.1) [R]. https://cran.r-project.org/web/packages/statConfR/index.html

Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, and Psychophysics*, *80*(1), 134–154. https://doi.org/10.3758/s13414-017-1431-5

Rausch, M., Hellmann, S., & Zehetleitner, M. (2021). Modelling visibility judgments using models of decision confidence. *Attention, Perception & Psychophysics*, *83*, 3311–3336. https://doi.org/10.3758/s13414-021-02284-3

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. https://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. https://doi.org/10.1371/journal.pcbi.1007456

Rausch, M., & Zehetleitner, M. (2023). Evaluating false positive rates of standard and hierarchical measures of metacognitive accuracy. *Metacognition and Learning*. https://doi.org/10.1007/s11409-023-09353-y

Rausch, M., Zehetleitner, M., Steinhauser, M., & Maier, M. E. (2020). Cognitive modelling reveals distinct electrophysiological markers of decision confidence and error monitoring. *NeuroImage*, *218*(116963), 1–14. https://doi.org/10.1016/j.neuroimage.2020.116963

Reynolds, A., Kvam, P. D., Osth, A. F., & Heathcote, A. (2020). Correlated racing evidence accumulator models. *Journal of Mathematical Psychology*, *96*, 102331. https://doi.org/10.1016/j.jmp.2020.102331

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. https://doi.org/10.1016/j.biopsych.2017.12.017

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience*, *1*(3), 165–175. https://doi.org/10.1080/17588921003632529

Said, N., Fischer, H., & Anders, G. (2022). Contested science: Individuals with higher metacognitive insight into interpretation of evidence are less likely to polarize. *Psychonomic Bulletin & Review*, *29*(2), 668–680. https://doi.org/10.3758/s13423-021-01993-y

Samaha, J., Barrett, J. J., Sheldon, A. D., LaRocque, J. J., & Postle, B. R. (2016). Dissociating perceptual confidence from discrimination accuracy reveals no influence of metacognitive awareness on working memory. *Frontiers in Psychology*, *7*(851), 1–8. https://doi.org/10.3389/fpsyg.2016.00851

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize ? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schwarz, G. (1978). Estimating the dimensions of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176348654

Shekhar, M., & Rahnev, D. (2021). The Nature of Metacognitive Inefficiency in Perceptual Decision Making. *Psychological Review*, *128*(1), 45–70. https://doi.org/10.1037/rev0000249

Shekhar, M., & Rahnev, D. (2022). How do humans give confidence? A comprehensive comparison of process models of metacognition. *PsyArXiv*. https://doi.org/10.31234/osf.io/cwrnt

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401–409. https://doi.org/10.1037/h0058700

Taouki, I., Lallier, M., & Soto, D. (2022). The role of metacognition in monitoring performance and regulating learning in early readers. *Metacognition and Learning*, 58. https://doi.org/10.1007/s11409-022-09292-0

Vandenbroucke, A. R. E., Sligte, I. G., Barrett, A. B., Seth, A. K., Fahrenfort, J. J., & Lamme, V. A. F. (2014). Accurate Metacognition for Visual Sensory Memory Representations. *Psychological Science*, *25*(4), 861–873. https://doi.org/10.1177/0956797613516146

Vlassova, A., Donkin, C., & Pearson, J. (2014). Unconscious information changes decision accuracy but not confidence. *Proceedings of the National Academy of Sciences*, *111*(45), 16214–16218. https://doi.org/10.1073/pnas.1403619111

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press.

Yates, J. F. (1990). *Judgment and decision making*. Prentice Hall.

Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of Postdecisional Processing of Confidence. *Journal of Experimental Psychology: General*, *144*(2), 489–510. https://doi.org/10.1037/xge0000062

Zhu, J.-Q., Sundh, J., Spicer, J., Chater, N., & Sanborn, A. N. (2023). The autocorrelated Bayesian sampler: A rational process for probability judgments, estimates, confidence intervals, choices, confidence judgments, and response times. *Psychological Review*. https://doi.org/10.1037/rev0000427

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in integrative Neuroscience*, *6*(79), 1–10. https://doi.org/10.3389/fnint.2012.00079