On the assumptions behind metacognitive measurements: Implications for theory and practice

Kiyofumi Miyoshi Graduate School of Informatics, Kyoto University, Kyoto, Japan

Yosuke Sakamoto Faculty of Letters, Kyoto University, Kyoto, Japan

Shin'ya Nishida Graduate School of Informatics, Kyoto University, Kyoto, Japan
NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Kanagawa, Japan

Theories of visual confidence have largely been grounded in the gaussian signal detection framework. This framework is so dominant that idiosyncratic consequences from this distributional assumption have remained unappreciated. This article reports systematic comparisons of the gaussian signal detection framework to its logistic counterpart in the measurement of metacognitive accuracy. Because of the difference in their distribution kurtosis, these frameworks are found to provide different perspectives regarding the efficiency of confidence rating relative to objective decision (the logistic model intrinsically gives greater meta-d′/d′ ratio than the gaussian model). These frameworks can also provide opposing conclusions regarding the metacognitive inefficiency along the internal evidence continuum (whether meta-d′ is larger or smaller for higher levels of confidence). Previous theories developed on these lines of analysis may need to be revisited as the gaussian and logistic metacognitive models received somewhat equivalent support in our quantitative model comparisons. Despite these discrepancies, however, we found that across-condition or across-participant comparisons of metacognitive measures are relatively robust against the distributional assumptions, which provides much assurance to conventional research practice. We hope this article promotes the awareness for the significance of hidden modeling assumptions, contributing to the cumulative development of the relevant field.

Introduction

Confidence rating has long been in use since the foundation of psychophysics (e.g., Egan & Clarke, 1956; Swets, Tanner, & Birdsall, 1955), and a currently predominant view is that confidence represents a metacognitive estimate of one's own decision correctness (e.g., Clarke, Birdsall, & Tanner, 1959; Galvin, Podd, Drga, & Whitmore, 2003). Since the advent of this research perspective, much interest has been paid to the matter of decision confidence in various research domains, including perception, memory, learning, consciousness, social interaction, and clinical applications (e.g., Guggenmos, Wilbertz, Hebart, & Sterzer, 2016; Peters, 2021; Rollwage & Fleming, 2021; Rouy, Saliou, Nalborczyk, Pereira, Roux, & Faivre, 2021; Toscani, Mamassian, & Valsecchi, 2021; Wixted & Wells, 2017).

Studies on decision confidence have been grounded in a variety of mathematical models. These models prescribe theoretically principled explanations of confidence rating behavior, helping researchers infer underlying operations for confidence construction (e.g., Denison, Adler, Carrasco, & Ma, 2018; Kiani, Corthell, & Shadlen, 2014; Ratcliff & Starns, 2013). One of the most fundamental frameworks for modeling confidence rating behavior is signal detection theory (SDT) (e.g., Green & Swets, 1966; Macmillan & Creelman, 2005). SDT posits that a certain degree of internal evidence occurs in response to perceptual input, which is collated to internal decision criteria to give rise to behavioral outputs (i.e., perceptual decision and confidence rating). Importantly, because of the existence of internal noise, even repeated presentations of the same stimulus are assumed to yield a variable degree of internal evidence from time to time, which follows certain forms of probability distribution (e.g., gaussian, logistic, Poisson, etc.). Here, the modeling choice of distributional forms involves some arbitrariness, and models of different distributional assumptions predict different shapes of receiver operating characteristics (ROCs). Thus one is advised to be careful that conclusions of certain

Citation: Miyoshi, K., Sakamoto, Y., & Nishida, S. (2022). On the assumptions behind metacognitive measurements: Implications for theory and practice. *Journal of Vision*, *22*(10):18, 1–15, https://doi.org/10.1167/jov.22.10.18.

https://doi.org/10.1167/jov.22.10.18 Received December 7, 2021; published September 23, 2022 ISSN 1534-7362

model-based analyses are robust against different distributional assumptions (e.g., Falmagne, 1985; Kellen, Winiger, Dunn, & Singmann, 2021).

Recent developments of SDT prescribe theoretically principled methods for evaluating the observer's metacognitive accuracy, which reflects the degree to which confidence ratings are diagnostic for decision correctness (Fleming & Lau, 2014; Maniscalco & Lau, 2012; Maniscalco & Lau, 2014). Metacognitive accuracy is estimated based on a so-called type 2 hit rate (type 2 HR, proportion of high confidence responses on correct decisions) and a type 2 false alarm rate (type 2 FAR, proportion of high confidence responses on incorrect decisions). Hereafter, we use the term type 2 decision referring to metacognitive decision on one's own decision correctness (e.g., confidence rating, point wagering, opt-out decision, etc.), and type 1 decision to mention the objective decision regarding external world states (e.g., Yes/No decision, multiple-alternative forced choice, etc.).

The cumulative type 2 HRs and FARs can be summarized in type 2 ROC space. The type 2 ROC demonstrates how well confidence rating discriminates between correct and incorrect type 1 responses, representing the observer's metacognitive accuracy. Furthermore, under some standard SDT assumptions, model fitting to type 2 ROC data can provide an estimate of metacognitive accuracy in terms of the normalized distance of two internal distributions that would have led to the observed type 2 ROC. This is a model-based measure of metacognitive accuracy, known as meta-d′ (Maniscalco & Lau, 2012, 2014). As well known, SDT also gives a parametric measure of type 1 accuracy termed d′. These two measures are known to be directly comparable, and thus studies have often used performance indices such as meta-d′ − d′ and meta-d′/d′ (hereafter called "m-ratio") to contrast observer's type 1 versus type 2 information processing efficiency. For example, the pattern of meta-d′ = d′ would be observed if both type 1 and type 2 decisions are made based on the same decision variable prescribed by a classic SDT model.

Metacognitive inefficiency is a default operation?

Contemporary theoretical understandings of visual metacognition have largely been shaped through the comparisons of meta-d′ against d′. Since the advent of the meta-d′ measure, studies have formed a rough consensus that empirically observed meta-d′ tends to be smaller than d′, a pattern of what we can call hypo-metacognitive sensitivity (Fleming & Daw, 2017; Maniscalco & Lau, 2012; Shekhar & Rahnev, 2020). Based on this observation, researchers have reached a theoretical view that internal evidence used for type 1 decision is further disrupted in type 2 calculation, or that the type 2 decision system does not have full access to the information used in type 1 decision. Either of the scenarios signifies the information loss at the type 2 processing level, and the multitude of components leading to this metacognitive inefficiency is theorized into a psychological construct known as metacognitive noise (e.g., Shekhar & Rahnev, 2020). Until now, metacognitive inefficiency relative to type 1 processing (m-ratio < 1) has often been supposed as a default hypothesis, whereas the pattern of m-ratio > 1 (known as hyper-metacognitive sensitivity) is considered an exceptional scenario that requires special considerations (e.g., Fleming & Daw, 2017; Shekhar & Rahnev, 2021).

Here, however, we would like to emphasize that this theoretical view has been built upon a particular distributional assumption; a predominant proportion of past studies have employed the gaussian SDT framework. One justification for the gaussian assumption comes from the central limit theorem. The gaussian assumption can be introduced, if not precisely, by thinking of internal sensory events composed of a multitude of independent subevents (Green & Swets, 1966, chapter 3). The gaussian distribution can also be introduced in relation to the maximum entropy principle (e.g., Norwich, 1993, chapter 8). Although these provide some credit to this framework, it is known that many forms of ROCs that are consistent with the gaussian distribution are also well captured by other probability distributions, and thus the gaussian assumption is ultimately unprovable (see Rouder, Pratte, & Morey, 2010). Therefore studies have advised against excessive dependence on particular modeling assumptions (Falmagne, 1985; Kellen & Klauer, 2014; Kellen & Klauer, 2015; Kellen et al., 2021).

These backgrounds have motivated us to systematically evaluate the behavior of the gaussian SDT in comparison to another influential decision-making framework, the logistic SDT. The logistic distribution has a firm theoretical basis in decision science as it is derived from the famous Luce's choice axiom (Luce, 1959, pp. 38–42; Macmillan & Creelman, 2005, pp. 94–104).[1] The logistic SDT has been actively in use for psychophysics (e.g., DeCarlo, 1998; Kornbrot, 2006; Pleskac, 2015), and there has been an emerging interest in its application to type 2 ROC analyses (Kristensen, Sandberg, & Bibby, 2020). Earlier literature stated that behaviors of the gaussian and logistic SDTs are practically indistinguishable (Luce, 1959), and they are unlikely to support discrepant conclusions in empirical analyses (Macmillan & Creelman, 2005). However, as we shall demonstrate later, the difference in their distributional forms has a considerable impact on the relative magnitude of meta-d′ against d′. In an extreme case scenario, the gaussian and logistic SDTs even provide qualitatively opposite conclusions regarding the observer's metacognitive efficiency (i.e.,

m-ratio > 1 or < 1), which would have significant implications for the theories of metacognition.

ZROC curvilinearity and criterion-dependency of metacognitive accuracy

Cumulative reports of gaussian m-ratio < 1 have led to the wide acceptance of the concept of metacognitive noise contamination. Then, studies have attempted to better characterize the underlying processes that cause metacognitive inefficiency (Samaha, Iemi, & Postle, 2017; Shekhar & Rahnev, 2020; Shekhar & Rahnev, 2021; Spence, Mattingley, & Dux, 2018; Zylberberg, Roelfsema, & Sigman, 2014). Shekhar and Rahnev (2021) have shown that empirical z-transformed type 1 ROCs consistently exhibited downward curvilinearity, a trend that cannot be captured by the standard gaussian SDT model (for zROC analysis in general, see Kellen & Klauer, 2018). They considered this trend as an indication of greater metacognitive inefficiency (lower metacognitive accuracy) toward higher confidence levels and proposed a mechanistic model that incorporates log-normally distributed metacognitive noise.

Here, again, we emphasize that their theoretical perspective postulates the gaussian SDT framework (z-transformation already implies gaussian underlying distributions). Downward curvature in zROC could appear without the incorporation of metacognitive noise if non-gaussian distributions are assumed at the type 1 level. In fact, this gave us a major motivation for considering the logistic distribution, as it has greater kurtosis (i.e., sharper peak and heavier tails) than the gaussian distribution, which, as we shall demonstrate later, naturally gives downward curvilinearity to zROC. This means that gaussian and logistic SDT analyses on the same dataset could reveal qualitatively discrepant conclusions on the operation of metacognitive inefficiency.[2]

These things considered, the present study compared the gaussian and logistic SDTs in their relevance to: (1) prevailing observations of m-ratio < 1 (i.e., metacognitive inefficiency relative to type 1 decision) and (2) zROC downward curvilinearity (i.e., metacognitive inefficiency along the internal evidence continuum). First, through computer simulations, we would systematically characterize the gaussian and logistic SDTs regarding their ROC predictions and parameter estimations. Then, we would compare these models in light of a large dataset (Rahnev, Desender, Lee, Adler, Aguilar-Lleyda, Akdoğan, Arbuzova, Atlas, Balcı, Bang, 2020) and evaluate the consequences resulting from the different distributional assumptions. Last, we would discuss how one could pursue constructive research practice based on the present findings.

Simulations with gaussian and logistic SDTs

In what follows, we use the term type 1 SDT models to refer to the traditional signal detection models that do not have a meta-d′ parameter (i.e., meta-d′ is constrained to be equal to d′). On the other hand, we refer to the models that have a meta-d′ parameter as meta-SDT models. Type-1 SDT models serve a baseline for comparison, showing default model behaviors without additional type 2 processes.

Figure 1 illustrates an example of the gaussian and logistic type 1 SDT models. The red and blue functions represent the distribution of internal evidence under two different external world states, S1 and S2. Upon an observation of an evidence sample, an observer makes a "S1" response if it falls left of the type 1 decision criterion (tick line in the middle) and responds "S2" otherwise.

We have simulated the gaussian and logistic type 1 SDT models, setting low, middle, and high performance conditions. Specifically, for the type 1 gaussian model, the difference of the means of S1 and S2 distributions was set at 1.00, 1.50, and 2.50 for the low, middle, and high performance conditions, whereas the standard deviation was set constant at 1 for all the conditions. For the type 1 logistic model, the difference of the means was set at 1.60, 2.44, and 4.28 for the three performance conditions, whereas the scale parameter was set constant at 1, which approximately corresponds to 1.81 in the unit of standard deviation. An unbiased type 1 decision criterion was set at the intersection of the two distributions for all the conditions. These parametrizations were determined so that the proportion of correct type 1 responses would be equated between the gaussian and logistic models at 0.69, 0.77, and 0.89 respectively for the low, middle, and high performance conditions; this ensured that the midpoint of the type 1 ROC (10th point from the left in Figure 2) for these models would be completely overlapped in each condition.

In each of the simulation conditions, we first simulated 400,000 type 1 decisions under the gaussian and logistic type 1 models. Then, we obtained 10 levels of confidence rating for each of those trials given S1 or S2 responses. In doing so, we used 10 quantiles of evidence samples as confidence criteria respectively for S1 and S2 responses. This procedure gives 19 cumulative data points in type 1 ROC space under each performance level.

Figure 2A shows the type 1 ROCs predicted by the type 1 SDT models. Colored dots demonstrate predictions of the type 1 logistic SDT and dashed lines show predictions of the type 1 gaussian SDT. Here, these ROCs can be interpreted as the observers' default performance characteristics under meta-d′ = d′. Distinct

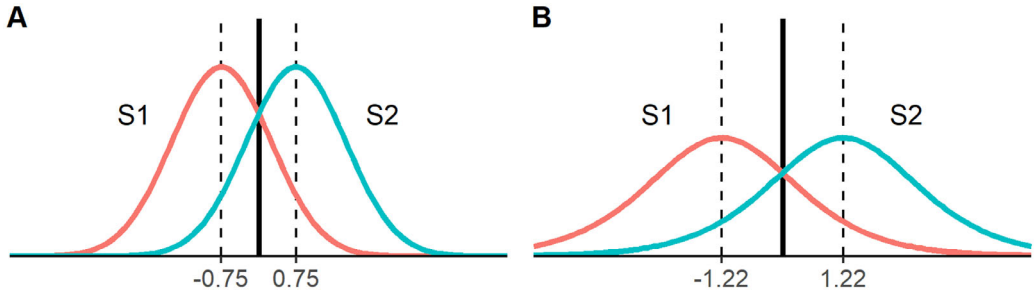

Figure 1. Type-1 gaussian SDT model (A) and type 1 logistic SDT model (B) under the middle performance condition (proportion of correct type 1 responses is matched between the models at 0.77). The models are depicted in their default scaling (standard deviation = 1 for the gaussian SDT and scale parameter = 1 for the logistic SDT).

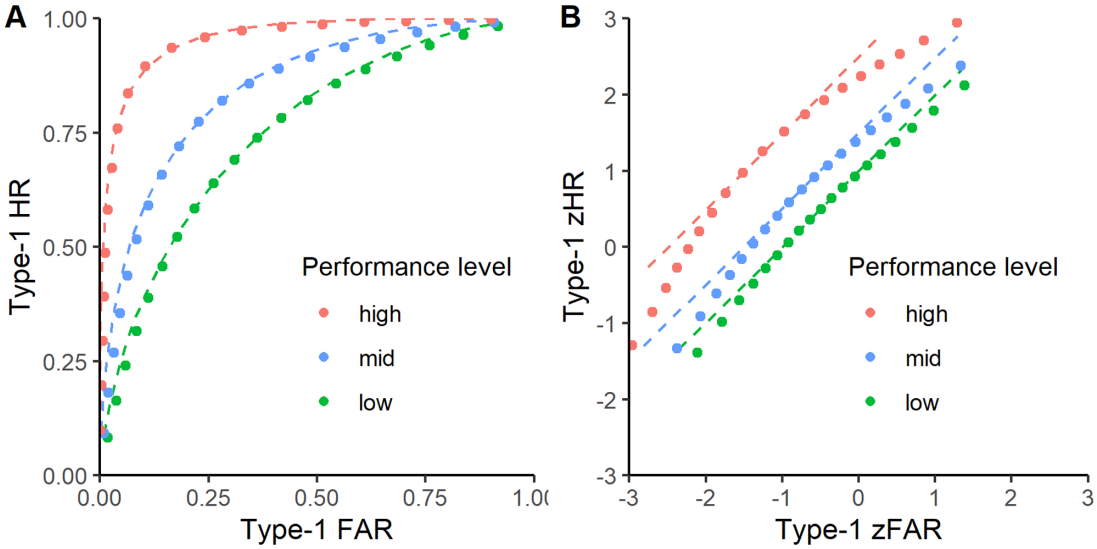

Figure 2. Type-1 ROCs derived from the type 1 gaussian SDT (*dashed lines*) and the type 1 logistic SDT (*dots*). The proportion of correct type 1 responses was matched between the models at each of the three performance levels. These curves demonstrate the models' default performance characteristics at m-ratio = 1.

ROCs can be seen between the models, reflecting the different shapes of the underlying distributions (logistic distribution has larger kurtosis). On a closer look, the gaussian type 1 ROCs are more protruded outward than those of the logistic type 1 SDT, even though we made the point 10th from the left, which is solely determined by the type 1 accuracy, completely overlapped between the models. Importantly, from the meta-SDT perspective, greater protruding curvature is indicative of greater metacognitive accuracy, whereas the middle data point is a pure representation of type 1 accuracy (e.g., type 1 ROC becomes a polygonal line concatenating the middle type 1 data point with [0, 0] and [1, 1] under zero metacognitive sensitivity). This means that compared to the logistic meta-SDT, the gaussian meta-SDT requires higher type 2 HR or lower type 2 FAR to yield the parametric pattern of meta-d′ = d′ under a certain type 1 accuracy. Accordingly, the pattern of m-ratio = 1 cannot be seen as the absolute benchmark for declaring metacognitive inefficiency as it is dependent on the auxiliary distributional assumption; the currently prevailing view that metacognition is not as informative as standard SDT prescribes could be underpinned by the somewhat arbitrary use of the gaussian modeling perspective.

The difference of the two models becomes more pronounced in zROC space (Figure 2B). By its definition, the gaussian type 1 SDT always gives linear zROC (e.g., Kellen & Klauer, 2018). Reflecting its distributional shape (sharp peak and heavy tails), the logistic type 1 SDT demonstrates distinctive downward curvilinearity, which becomes more salient as a function of the performance level. Again, the middle type 1 data point (10 from the left) was completely overlapped between the models. Yet, the other data points, which reflect the observers' metacognitive accuracy, fell lower

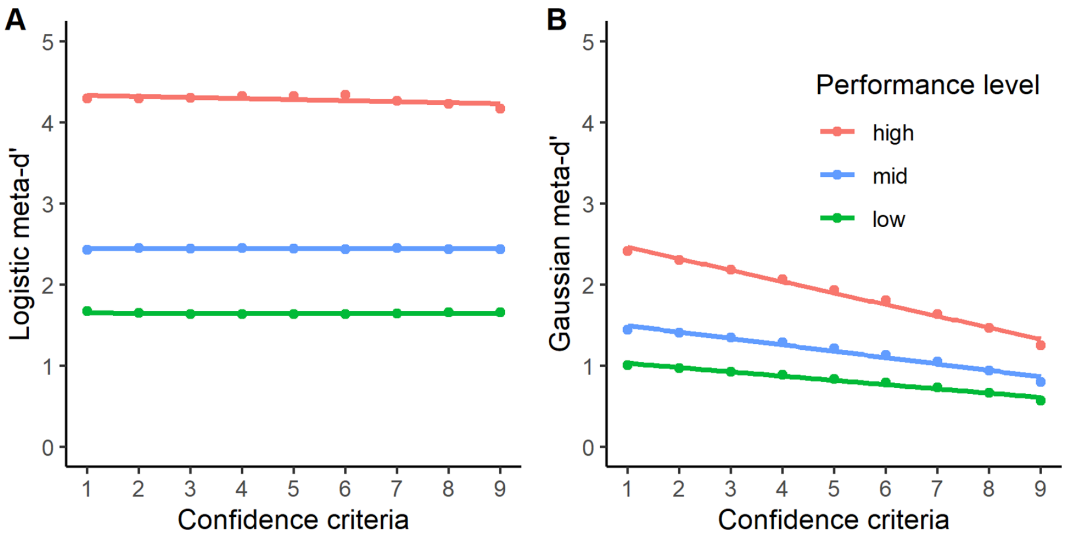

Figure 3. Meta-d′ estimated at different confidence criteria under the logistic meta-SDT (A) and the gaussian meta-SDT (B). The horizontal axis shows confidence criteria ordered from lowest to highest. Logistic estimates are shown at approximately 1.81 times the scale than gaussian estimates.

	d′	Meta-d′	m-ratio
Low condition			
Gaussian meta-SDT	0.993	0.754	0.759
Logistic meta-SDT	1.602	1.590	0.992
Middle condition			
Gaussian meta-SDT	1.486	1.160	0.780
Logistic meta-SDT	2.431	2.431	1.00
High condition			
Gaussian meta-SDT	2.505	2.103	0.839
Logistic meta-SDT	4.282	4.316	1.008

Table 1. Estimates of the meta-SDT models based on the logistic simulation data. Logistic estimates are shown at approximately 1.81 times the scale than gaussian estimates.

for the logistic than gaussian models (i.e., the logistic SDT framework allows lower type 2 HR and/or greater type 2 FAR than the gaussian SDT to demonstrate the parametric pattern of m-ratio = 1).

Shekhar and Rahnev (2021) found downward curvature in empirical zROCs, and, under the gaussian SDT perspective, proposed the view that greater metacognitive noise occurs toward the ends of the internal evidence continuum. However, the current simulation demonstrates that modeling perspectives other than the gaussian SDT could naturally explain the zROC curvilinearity without incorporating extra metacognitive noise components. Furthermore, it is even possible that analyses based on non-gaussian perspectives could reveal qualitatively different criterion-dependency of metacognitive accuracy than the one proposed in the previous literature.

To ascertain these insights, we have fitted the gaussian and logistic meta-SDT models to the aforementioned simulation data sampled from the type 1 logistic SDT (Table 1). As expected, the logistic meta-SDT showed a larger m-ratio than the gaussian meta-SDT, retrieving the original constraint of the type 1 logistic SDT (meta-d′ = d′). The gaussian meta-SDT estimated meta-d′ to be smaller than d′ even though no type 2 disruption was implemented in the logistic simulation data. Furthermore, we have evaluated metacognitive accuracy at each location of the nine confidence criteria (see later sections for methodological details). Estimates of the logistic meta-SDT were invariant against the location of confidence criteria, constantly showing m-ratio around 1 (Figure 3A). However, the gaussian meta-SDT estimated metacognitive accuracy to become poorer for higher confidence levels (Figure 3B), suggesting as if there is growing contamination of metacognitive noise towards the ends of the evidence continuum. These simulation results demonstrate that there is much room for reconsidering the hitherto established understanding of visual metacognition from alternative perspectives than the gaussian SDT, motivating the following large-scale analyses.

Analyses on the confidence database

We have fitted the gaussian and logistic SDT models to large datasets from the confidence database (Rahnev et al., 2020), whereby we aimed to evaluate the generality

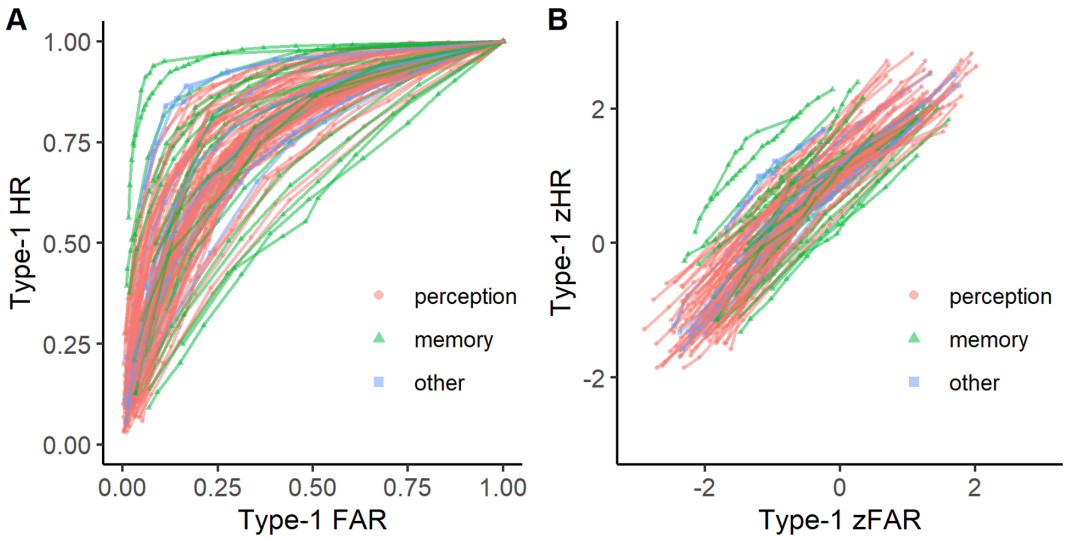

Figure 4. Empirical type 1 ROCs in original space (A) and z-transformed space (B). Data from individual participants were aggregated for each of the 105 cases. The color indicates the cognitive domain to which each case belongs.

of the knowledge that empirical m-ratio tends to be less than 1 and that metacognition becomes worse at higher confidence criteria. We have also been interested in quantitative model comparisons. The SDT models considered here are supposed to be measurement tools rather than precise process models of metacognition. Thus, our model comparisons did not aim to determine whether the true underlying distribution is gaussian or logistic. Rather, the model comparisons were made to ascertain the validity of the model-based measurements (i.e., measurements made on a poorly fitting model are considered to be invalid) (Myung & Pitt, 2018).

Following methodological considerations from past studies (Lee, Ruby, Giles, & Lau, 2018; Maniscalco & Lau, 2014), we have only included genuine two-alternative forced-choice (2AFC) data in the present analyses.[3] Here, the genuine 2AFC refers to the task design which offers two explicit stimulus intervals on every trial and requires participants to identify which of the two includes the target stimulus (Macmillan & Creelman, 2005, chap. 7). In order to evaluate metacognitive accuracy's criterion-dependency, we have only targeted experiments that employed four or more levels of confidence rating. The following analyses were conducted on the free statistical language R (Version 4.0.5).

For those studies that used a continuous confidence scale, we have used a round() function on R to obtain ordinal confidence rating data; this function transforms continuous values to the nearest integer according to the IEC 60559 standard. For example, continuous values distributed between 0 and 10 were converted into 11-step integer values, which we have counted as an ordinal confidence scale of 11 levels. When necessary, an original continuous scale was divided or multiplied by a multiple of 10 before the round() function was applied (e.g., continuous values ranging from 50 to 100 were first divided by 10 and rounded into six-step integer values [5, 6, 7, 8, 9, 10], which was deemed an ordinal confidence scale of six levels [see Supplementary Material 1]).

For the datasets that include categorical experimental manipulations (e.g., low vs. high difficulty), we have counted each condition as a separate case and independently fit the SDT models. In the cases of trial-by-trial continuous manipulations (e.g., jittered contrast across trials), we have fitted the models with aggregating the trials over different stimulus intensities (see Supplementary Materials 1 and 2 for further details).

Based on the data selection procedure, we have identified 105 relevant cases of 2AFC experimental conditions (70 from perception, 19 from memory, and 16 from other cognitive domains), which include 5160 cases at the individual participant level (Supplementary Material 1). Figure 4 shows empirical type 1 ROCs for the 105 cases, where responses from individual participants are aggregated in each case. As is usual in 2AFC data, these type 1 ROCs are highly symmetrical about the major diagonal, demonstrating suitability for the meta-SDT analysis. Importantly, many of the ROCs indicate downward curvature in the z-transformed space, replicating the key observation from Shekhar and Rahnev (2021); quadratic regressions revealed 73 cases of significant downward curvature, 11 cases of significant upward curvature, and 21 cases of insignificance.

In what follows, we fit four different SDT models to individual participants' data. These models differ in the distributional assumption (gaussian vs. logistic), as well

as the presence/absence of the meta-d′ parameter (type 1 SDT vs. meta-SDT). Before numerical model fittings, we first calculated type 1 criterion (θ) and sensitivity (d′) according to the following equations (Kristensen et al., 2020), where HR1 and FAR1 are empirical type 1 hit and false alarm rates, and z() represents a transformation with the inverse gaussian cumulative distribution function.

$$\theta_{gaussian} = -z(FAR1)$$

$$d'_{gaussian} = z(HR1) - z(FAR1)$$

$$\theta_{logistic} = -log\left(\frac{FAR1}{1-FAR1}\right)$$

$$d'_{logistic} = log\left(\frac{HR1}{1-HR1}\right) - log\left(\frac{FAR1}{1-FAR1}\right)$$

Then, we fit these models to type 2 ROC data by the maximum likelihood method while constraining the relative type 1 criterion (Θ) as follows (see Barrett, Dienes, & Seth, 2013; Maniscalco & Lau, 2014):

$$\Theta_{gaussian} = \frac{\theta_{gaussian}}{d'_{gaussian}}$$

$$\Theta_{logistic} = \frac{\theta_{logistic}}{d'_{logistic}}$$

Meta-d′ and confidence criteria were estimated in the meta-SDT model fitting. On the contrary, the type 1 SDT model fitting only estimated confidence criteria under the constraint of meta-d′ = d′, which corresponds to assuming that type 1 and type 2 decisions are made on the same decision variable (i.e., no contamination of metacognitive noise or acquisition of additional metacognitive evidence). Thus the type 1 SDT models serve as null models to determine whether one needs the extra meta-d′ parameter to sufficiently explain participants' metacognitive behavior. Namely, by comparing the type 1 and meta-SDT models, one can test the presence of metacognitive inefficiency without specifying the exact mechanism underlying its occurrence. Because the type 1 parameters were analytically calculated through the above equations, goodness-of-fit indexes reported below are only pertaining to type 2 ROC fittings.

In the fitting of type 2 data, one needs to consider two sets of equations: one for type 2 performances for S1 responses and the other for S2 responses. Under the gaussian SDT models, the S1-response-specific type 2 hit rate ($hr2_{S1}$) and the S1-response-specific type 2 false alarm rate ($far2_{S1}$) were defined by the following equations, where $\tau_{gaussian\_S1}$ is the confidence criterion on the S1 response side, meta-$d'_{gaussian}$ is the gaussian meta-d′, Φ_0 is the cumulative standard gaussian distribution function, and $\Phi_{meta\text{-}d'\_gaussian}$ is the cumulative gaussian distribution function with the standard deviation of 1 and the mean value that is equal to the gaussian meta-d′. In the fitting of multilevel confidence rating data, $\tau_{gaussian\_S1}$ becomes a vector containing a series of confidence criteria, and $hr2_{S1}$ and $far2_{S1}$ are defined at each location of those.

$$hr2_{S1} = \frac{\Phi_0\left(\tau_{gaussian\_S1}\right)}{\Phi_0\left(meta - d'_{gaussian} \times \Theta_{gaussian}\right)}$$

$$far2_{S1} = \frac{\Phi_{meta-d'\_gaussian}\left(\tau_{gaussian\_S1}\right)}{\Phi_{meta-d'\_gaussian}\left(meta - d'_{gaussian} \times \Theta_{gaussian}\right)}$$

The S2-response-specific type 2 hit rate ($hr2_{S2}$) and the S2-response-specific type 2 false alarm rate ($far2_{S2}$) were defined as follows, where $\tau_{gaussian\_S2}$ is the confidence criterion (or criteria) on the S2 response side.

$$hr2_{S2} = \frac{1 - \Phi_{meta-d'\_gaussian}\left(\tau_{gaussian\_S2}\right)}{1 - \Phi_{meta-d'\_gaussian}\left(meta - d'_{gaussian} \times \Theta_{gaussian}\right)}$$

$$far2_{S2} = \frac{1 - \Phi_0\left(\tau_{gaussian\_S2}\right)}{1 - \Phi_0\left(meta - d'_{gaussian} \times \Theta_{gaussian}\right)}$$

These four equations collectively provide the estimated probability for all the response categories; 2AFC data with n levels of confidence are classified into 2 (S1 or S2 responses) × 2 (type 2 hit or type 2 false alarm) × n (confidence levels) response categories (e.g., S1-response-specific type 2 hit rate with the confidence level of 3). We have estimated the parameters by maximizing the log-likelihood over all the response categories, which is defined in the following equation. Here, n, r, and t refer to the possible values of confidence level, response class (S1 or S2), and outcome type (hit or false alarm). Consequently, this equation represents the sum of the products between the logarithm of the estimated probability of each response category and the observed response frequency for the corresponding category (for further technical details, see Barrett et al., 2013; Maniscalco & Lau, 2014).

$$LL = \prod_{n,\ r,\ t} \log\left(p\left(Conf_n \mid Resp_r, Type_t\right)\right) \times Freq\left(Conf_n \mid Resp_r, Type_t\right)$$

Under the logistic models, the response-specific type 2 hit and type 2 false alarm rates were defined as follows, where meta-$d'_{logistic}$ is the logistic meta-d′, Λ_0 is the cumulative logistic distribution function with the scale parameter of 1 and the location parameter of 0, while

	d′	Meta-d′	m-matio	Case converged	Case above-chance
Gaussian meta-SDT	1.390	1.219	0.948	4056/5160	3818/4056
Logistic meta-SDT	2.346	2.269	1.052	4355/5160	4127/4355
Gaussian type 1 SDT	1.370	1.370	1	4208/5160	4146/4208
Logistic type 1 SDT	2.313	2.313	1	4442/5160	4376/4442

Table 2. Model fits to individual data. Performance measures were averaged across the cases where each model converged and showed above-chance type 1 and type 2 performances. Logistic estimates are shown at approximately 1.81 times the scale than gaussian estimates.

$\Lambda_{\text{meta-d'\_logistic}}$ is the cumulative logistic distribution function with the scale parameter of 1 and the location parameter that is equal to the logistic meta-d′. Also, $\tau_{\text{logistic\_S1}}$ and $\tau_{\text{logistic\_S2}}$ are the confidence criterion (or criteria) on each response side. As in the case of the gaussian fitting, the parameters were estimated by the maximization of the log-likelihood, which is again given by the sum of the products of the log-estimated probability of each response category and the observed response frequency for the corresponding category.

$$hr2_{S1} = \frac{\Lambda_0\left(\tau_{logistic\_S1}\right)}{\Lambda_0\left(meta-d'_{logistic} \times \Theta_{logistic}\right)}$$

$$far2_{S1} = \frac{\Lambda_{meta-d'\_logistic}\left(\tau_{logistic\_S1}\right)}{\Lambda_{meta-d'\_logistic}\left(meta-d'_{logistic} \times \Theta_{logistic}\right)}$$

$$hr2_{S2} = \frac{1-\Lambda_{meta-d'\_logistic}\left(\tau_{logistic\_S2}\right)}{1-\Lambda_{meta-d'\_logistic}\left(meta-d'_{logistic} \times \Theta_{logistic}\right)}$$

$$far2_{S2} = \frac{1-\Lambda_0\left(\tau_{logistic\_S2}\right)}{1-\Lambda_0\left(meta-d'_{logistic} \times \Theta_{logistic}\right)}$$

For model comparisons, we have used Akaike information criterion (AIC) and Bayesian information criterion (BIC) defined as follows, where log-likelihood denotes the log-likelihood of the model, K is the number of free parameters, and N is the total number of trials across all response categories. Note that the type 1 SDT models, whose meta-d′ was fixed to be equal to d′, have one less free parameter than the meta-SDT models.

$$AIC = -2LL + 2K$$

$$BIC = -2LL + log(N)\,K$$

Model fits to individual data

Table 2 summarizes the fittings to individual participants' data (Supplementary Material 2 also reports fittings to aggregated data). As is usual in individual fittings, there were quite a few cases where the estimation did not converge. The logistic models showed some advantage over the gaussian models in terms of the number of converged cases. Because the number of converged cases differed across the models, we could not simply calculate summed information criteria over individual cases for model comparisons. Thus, across the total of 5160 individual cases, we examined the number of cases for which each model declared the best AIC/BIC fit. Models that failed to converge were disqualified for each case, and no winner was declared in those cases where all four models did not converge.

	Best AIC fits	Best BIC fits
Gaussian meta-SDT	871	438
Logistic meta-SDT	826	438
Gaussian type 1 SDT	1409	1770
Logistic type 1 SDT	1672	2132

Table 3. Number of the cases for which each of the models showed best fit in terms of AIC and BIC.

The analysis adjudicated the type 1 logistic model as a clear victor, presumably because it naturally captures curvilinearity in zROCs without incorporating the extra meta-d′ parameter (Table 3). The type 1 SDT models were generally favored over the meta-SDT models, suggesting that the occurrence of metacognitive inefficiency may not be taken as much for granted as previously considered. This means that some of those previous reports from the gaussian perspective, originally constituting supporting evidence for metacognitive inefficiency, could be better explained by the type 1 logistic model postulating flawless metacognition (m-ratio = 1). However, it should be noted that the inclusion of extra parameters is difficult to be justified in individual analyses due to the matter of statistical power. There was no clear-cut winner in the comparison of the gaussian and logistic meta-SDT models, indicating that the logistic meta-SDT is no less viable than the gaussian meta-SDT as a measurement model of metacognitive accuracy.

Next, we would evaluate parameter estimates on those cases in which each of the models converged and revealed above-chance type 1 and type 2 performances

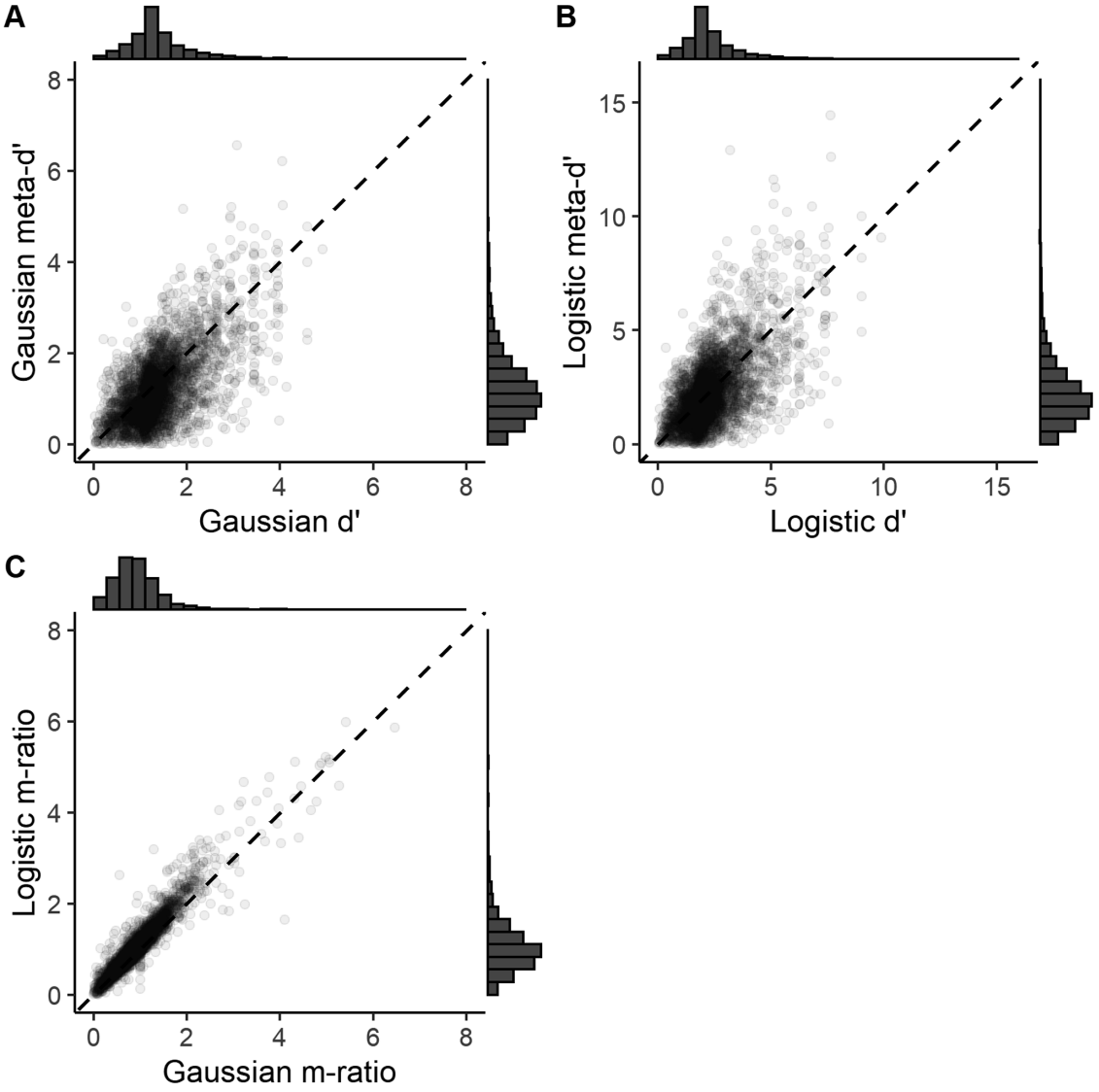

Figure 5. Parameter estimates of the gaussian and logistic meta-SDTs. Cases failed to reach chance performances were excluded here, which left 3818 cases for the gaussian meta-SDT (A) and 4127 cases for the logistic meta-SDT (B). Panel C only includes 3552 cases for which both models showed above-chance performances. Logistic estimates are shown at approximately 1.81 times the scale than gaussian estimates.

(Table 2). Paired *t*-tests showed that mean meta-d′ was significantly smaller than mean d′ under the gaussian meta-SDT (t (3817) $= -16.99$, $p < 0.001$, Figure 5A) and the logistic meta-SDT (t (4126) $= -4.03$, $p < 0.001$, Figure 5B), indicating metacognitive inefficiency on average basis. However, caution would be advised because the logistic meta-SDT indicated greater mean m-ratio than the gaussian meta-SDT ($t = 28.09$, $p < 0.001$, Figure 5C), and there are even cases where the gaussian meta-SDT showed m-ratio < 1 whereas the logistic meta-SDT demonstrated m-ratio > 1 (399 of 3552 cases in Figure 5C). These results exemplify the important consequences from the distributional assumptions. In an extreme scenario, it is even possible that the models advocate for the qualitatively opposite theoretical operations (i.e., contamination of metacognitive noise vs. acquisition of metacognitive evidence).

We found rather frequent occurrences of hyper-metacognitive sensitivity (m-ratio > 1) in Figure 5. A Fisher's exact test showed that this observation was more frequent under the logistic than gaussian meta-SDTs (1873 of 4127 vs. 1412 of 3818 cases, $p < 0.001$). Despite these differences, however, m-ratios estimated by these models were highly consistent across the individual cases (Pearson's $r = 0.942$), ensuring the models' reliability in the assessment of metacognitive accuracy (Figure 5C).

Lastly, we have examined criterion-dependency of metacognitive accuracy by estimating meta-d′

parameters at different confidence criteria. For this purpose, we have converted multilevel confidence rating data into binary formats (i.e., high vs. low confidence) by making dichotomous cutoffs at different confidence criteria (Rahnev, 2021; Shekhar & Rahnev, 2021). To illustrate this, let us consider a response frequency dataset of (2, 3, 5, 7, 11, 13), which is comprised of two type 1 response classes and three levels of confidence rating (i.e., ten S1 responses [sum of the first three] and thirty-one S2 responses [sum of the last three] are bounded by lower and higher confidence criteria, constituting a sequence of response frequencies from highest confidence S1 to highest confidence S2). This dataset will be (5, 5, 7, 24) by making a binary cutoff at the lower confidence criterion while cutting off at the higher confidence criterion gives (2, 8, 18, 13). Thus a dataset of n levels confidence rating allows for n-1 different binary cutoffs, and we have estimated meta-d′ for each of the reformatted datasets; speaking of the current example, meta-d′ estimated on the dataset (5, 5, 7, 24) represents metacognitive accuracy at the lower confidence criterion, whereas that on (2, 8, 18, 13) indicates metacognitive accuracy at the higher confidence criterion.

The binary reformatting yielded a total of 34975 datasets, and both the gaussian and logistic meta-SDTs converged in 30512 cases. Among those, we have included 26927 cases in the present analysis for which both the models exhibited above-chance type 1 and type 2 performances. Because different levels of confidence rating were employed across studies, we have normalized confidence criteria for the subsequent analysis. For example, regarding confidence data of 6 levels, we have estimated meta-d′ for 5 different binary reformatted

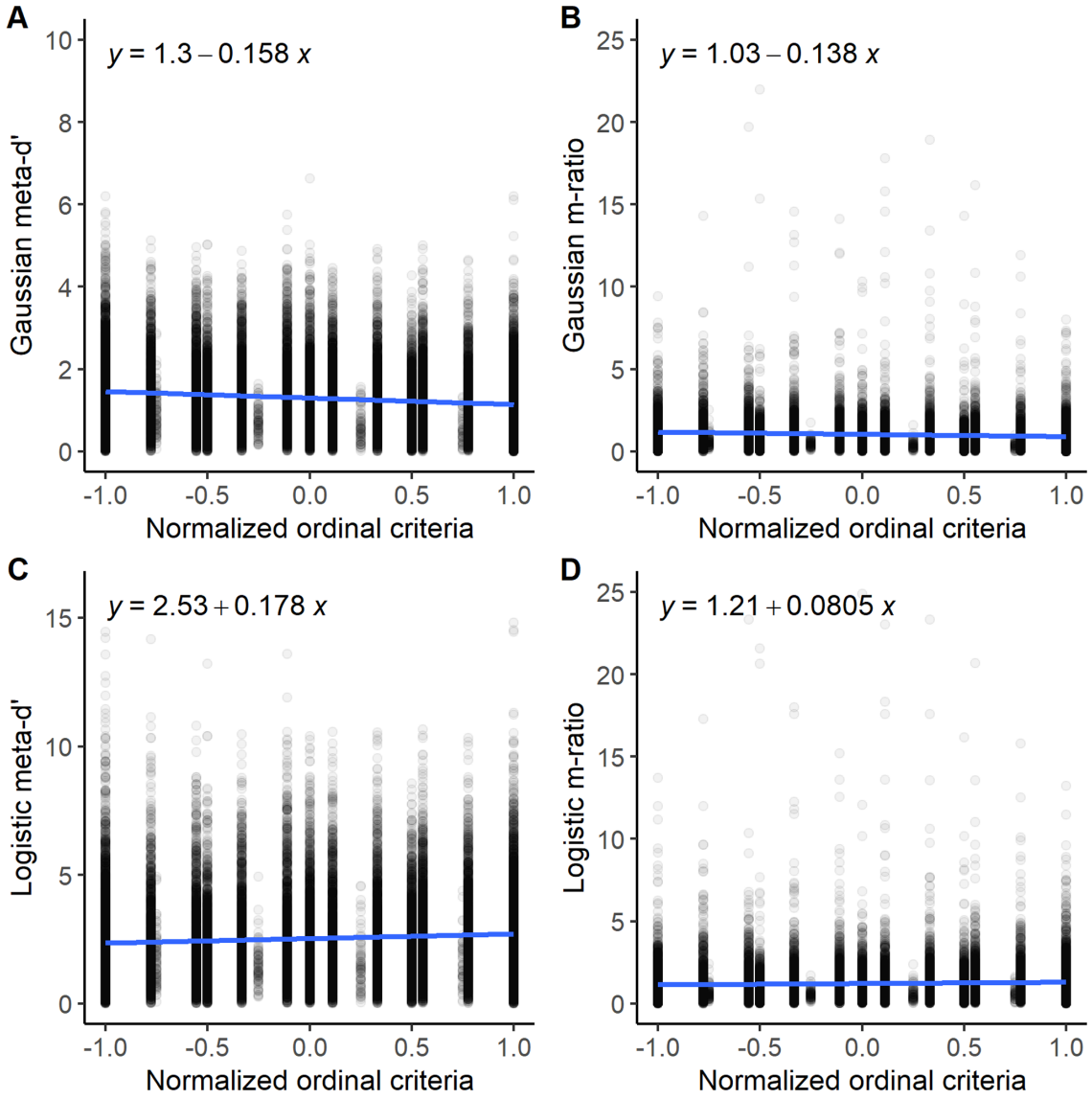

Figure 6. Metacognitive performances evaluated at different confidence criteria. The criteria are normalized and ordered from lowest to highest (lowest/highest criteria are coded as −1/1). Cases across different individuals and studies are shown all together. Logistic estimates are shown at approximately 1.81 times the scale than gaussian estimates.

data, and then confidence criteria associated with these meta-d′ values are numbered as [1, 2, 3, 4, 5] from lowest to highest. Next, we have normalized these ordinal numbers so that the lowest (and highest) criterion would be coded as −1 (and 1) with intermediate criteria evenly spaced between them (i.e., confidence criteria coded as [1, 2, 3, 4, 5] were normalized into [−1, −0.5, 0, 0.5, 1]). Last, metacognitive performances were regressed by the normalized ordinal criteria, which captures rather qualitative trends of criterion-dependency; the estimated slope is scaled to indicate the performance difference for middle and extreme levels of confidence.[4]

We conducted linear regression to explain metacognitive performances from the normalized ordinal criteria and tested if the slope is significantly different from 0 (26927 data points are aggregated in each panel of Figure 6). The results showed negative criterion-dependency for gaussian meta-d′ ($t = -3921.54$, $p < 0.001$) and gaussian m-ratio ($t = -16.52$, $p < 0.001$), which is consistent with the previous report of greater metacognitive inefficiency at higher confidence criteria (Rahnev, 2021; Shekhar & Rahnev, 2021). Importantly, however, the logistic meta-SDT showed positive criterion-dependency for meta-d′ ($t = 11.27$, $p < 0.001$) and m-ratio ($t = 38.20$, $p < 0.001$), indicating that metacognition is more efficient at higher confidence criteria. The reversal of the criterion dependency again showcases the serious consequence of the auxiliary modeling assumptions.

Discussion

We have demonstrated important discrepancies between the gaussian and logistic SDT frameworks in the assessment of metacognitive accuracy. There has been a widespread view that the human metacognitive system is unable to make full use of the information employed by the type 1 decision system (e.g., Maniscalco & Lau, 2012; Shekhar & Rahnev, 2020, 2021). Our analyses on the large-scale individual datasets partly confirmed this view since, on average, meta-d′ was estimated to be smaller than d′ under both the gaussian and logistic meta-SDTs. However, our results are somewhat mixed here as the model comparisons best favored the type 1 logistic SDT, which does not incorporate any metacognitive inefficiency. Of practical importance is that the logistic meta-SDT intrinsically gives greater m-ratio than the gaussian meta-SDT because of the difference in their distribution kurtosis. Therefore, in extreme cases, these models can advocate for opposing interpretations on the same dataset (gaussian m-ratio < 1 vs. logistic m-ratio > 1).

The present findings suggest that the observation of m-ratio ≥ 1 is not as unnatural as previously considered, alleviating the need to presuppose the existence of metacognitive inefficiency as a default hypothesis. In fact, there is no intrinsic reason within the realm of SDT that restricts m-ratio less than 1, and our previous studies have indeed demonstrated that the pattern of m-ratio ≥ 1 naturally emerges as a consequence of our metacognitive system being adapted to statistical structures of the world (see Miyoshi & Lau, 2020; Webb, Miyoshi, So, Rajananda, & Lau, 2021).

Notice, however, that we are not arguing that metacognitive inefficiency does not really exist or that the type 1 SDT is sufficient on its own to explain both type 1 and type 2 decisions. Rather, we see type 1 versus type 2 dissociation evidenced in those cases where both gaussian and logistic m-ratios are fairly smaller (or larger) than 1 (see Figure 5). The objective versus subjective dissociation is further supported by intervention studies demonstrating rather selective impairments of type 2 accuracy with transcranial magnetic stimulation (Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010), muscimol injection (Miyamoto, Osada, Setsuie, Takeda, Tamura, Adachi, & Miyashita, 2017), and secondary task demand (Maniscalco & Lau, 2015). Therefore our claim here is that it is not compulsory to put asymmetric preference on metacognitive inefficiency (m-ratio < 1) over efficiency (m-ratio > 1), and researchers can search for underlying causes of these observations from a neutral point of view.

Past studies have suggested several factors that may influence metacognitive efficiency. One thing is that if there is a time lag between type 1 and type 2 decisions (i.e., participants indicate their type 1 decision first and then rate their confidence), additional metacognitive evidence could be collected to boost the accuracy of metacognition (e.g., Fleming & Daw, 2017; Pleskac & Busemeyer, 2010). However, this effect was difficult to be examined in the present study because such two-stage response design was employed in 98 of the 105 targeted datasets. For another thing, greater metacognitive efficiency has been observed when trials of varying difficulties are intermixed in a single experimental block (Rahnev & Fleming, 2019), although we could not find favorable evidence for this effect at least in the present datasets (see Table S2 in Supplementary Material 2). In any case, the current datasets include a variety of uncontrolled features, making it difficult to identify factors that contribute to efficient metacognition. Numerous factors, including the above two, remain of interest for future controlled studies.

The criterion-dependency of metacognitive accuracy is another topic on which the gaussian and logistic SDT frameworks have shown disagreement. In an attempt to better characterize the nature of metacognitive efficiency, recent studies have reported that gaussian meta-d′ became smaller at higher confidence criteria (Rahnev, 2021; Shekhar & Rahnev,

2021). They developed a model that incorporates log-normally distributed metacognitive noise on top of gaussian type 1 distributions to describe growing metacognitive inefficiency towards the ends of the evidence continuum. However, we have shown that the criterion-dependency was reversed (i.e., greater metacognitive efficiency toward the ends of the continuum) if the underlying distribution is assumed to be logistic, instead of gaussian. This is mainly because the logistic distribution has greater kurtosis and predicts different operating characteristics than the gaussian distribution (Figure 2), which suggests that the criterion-dependency of metacognitive accuracy is rather difficult to be determined by means of behavioral modeling.

These findings showcase the fact that theories of metacognitive operations have been formed through certain forms of mathematical models standing upon auxiliary distributional assumptions. Thus, the reliability of a certain research conclusion should be assessed in regard to its robustness against underlying modeling assumptions (for model comparisons on minimal assumptions, see Kellen & Klauer, 2014; Kellen & Klauer, 2015; Miyoshi, Kuwahara, & Kawaguchi, 2018). There is no one-to-one relationship between a certain theoretical construct (e.g., internal evidence) and its possible implementations in mathematical models (e.g., sample from gaussian or logistic distributions). Accordingly, the same data could be equally well explained by different models that even provide qualitatively opposite theoretical interpretations.

One countermeasure we can take against this problem is to see agreement across different model variants, a method called "multiverse" analysis (e.g., Oberauer & Lewandowsky, 2019; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). One should be assured to insist on a certain research conclusion (e.g., metacognition is inefficient than objective type 1 decision) if different models provide consistent support for it. The comparison of multiple model variants may further provide some valuable insights into the cognitive process in question. For example, in the present case, meta-d′ estimated by the gaussian and logistic meta-SDTs exhibited the reversed criterion-dependency. This suggests that a distribution with a kurtosis intermediate of theirs may be an appropriate candidate for the underlying function of internal evidence (see Supplementary Figure S4 in Supplementary Material 2 for further details).

When evaluating the agreement of different models, one can weight them according to the extent that these models are supported by the data in question. We originally hoped that the current large-scale model comparisons would adjudicate between the gaussian and logistic meta-SDT models, but the results were rather equivocal in this regard. Perhaps, the curvilinearity in zROC, the qualitative trend that cannot be explained by the type 1 gaussian SDT model, is appreciably captured by introducing the meta-d′ parameter, which makes it difficult to decide a clear victor between the gaussian and logistic meta-SDT models. Of course, the validity of cognitive models is not solely determined by the goodness of fit to empirical data but is also evaluated by other standards such as consistency to existing domain knowledge (Myung & Pitt, 2018). Yet, the gaussian and logistic SDTs seem not to be easily distinguishable in this regard as they both have been of successful use in various domains of decision science (e.g., Macmillan & Creelman, 2005). Because no clear winner was found in the present model comparisons, it is difficult to prescribe a general recommendation regarding which model's estimates should be more prioritized in practical use. For the time being, it might be safer to avoid overreliance on either of the models.

Nevertheless, it is worth emphasizing that the m-ratio measures of the gaussian and logistic meta-SDTs exhibited high degrees of consistency when estimations were made on the full range of confidence levels (Figure 5C). This means that condition-wise comparisons (i.e., whether condition A or B leads to higher metacognitive efficiency) or across-participant correlation analyses (i.e., if participants who achieve great metacognitive efficiency in task A also tend to perform well in task B) are rather invariable against the choice between the measurement models. On the contrary, one needs to be cautious when directly comparing meta-d′ against d′ to see if m-ratio would be larger/smaller than 1, or when evaluating metacognitive accuracy at each location of confidence criteria. The outcomes of these analyses are quite sensitive to auxiliary distributional assumptions.

One thing to note is that the present study only selected genuine 2AFC data for use. This is mainly because data from Yes/No experiments usually demonstrate asymmetric type 1 ROC, which is not compatible with the meta-SDT analysis (from the SDT perspective, this asymmetry indicates unequal variance between target and nontarget distributions in unidimensional decision space). On the contrary, due to the symmetric treatment of S1 and S2 responses in bidimensional decision space, 2AFC experiments are supposed to provide symmetric type 1 ROC, regardless of the variance-covariance structure of target and nontarget distributions (for graphical intuitions, see Miyoshi & Lau, 2020). Also, importantly, Miyoshi and Lau (2020) has shown that certain metacognitive heuristics can lead to better metacognition under appropriate variance-covariance structure in 2AFC experiments, which we deem as an intrinsic property of our metacognitive system. Thus the measurements made for the present 2AFC datasets may reflect the mixture of different components (e.g., internal metacognitive noise, response noise of confidence

rating, appropriate use of metacognitive heuristics, etc.), the relative contributions of which should determine the observer's metacognitive efficiency (also see Supplementary Table S8 for supplementary analyses on non-2AFC data).

Although it may sound daunting, we believe it is an important step for us to acknowledge that there can be one-to-many correspondence between a theoretical operation of interest and its modeling implementations. We shall strive for establishing essential theories of visual metacognition that are not predicated on strong auxiliary assumptions. Through such practices, one can be more certain about her research conclusions and make solid contributions to the field's sustainable development.

Keywords: confidence, metacognition, metacognitive inefficiency, signal detection theory, theory development

Acknowledgments

The authors thank Hakwan Lau and Dobromir Rahnev for valuable input on this research.

Supported by JSPS in the form of Overseas Research Fellowship and KAKENHI Grant Number 22K13870 awarded to Kiyofumi Miyoshi.

Commercial relationships: none.
Corresponding author: Kiyofumi Miyoshi.
Email: miyoshi80@gmail.com.
Address: Graduate School of Informatics, Kyoto University, Sakyo, Kyoto 6068501, Japan.

Footnotes

[1]The way the logistic framework is grounded in the Luce axiom is different from the way the gaussian framework is supported by the central limit theorem. The central limit theorem provides a direct basis for assuming the gaussian distribution, whereas the Luce axiom is a set of assumptions that provide a starting point for interpreting the estimates of the logistic SDT. Thus some may think that the gaussian distribution has a stronger basis than the logistic distribution in psychophysics. However, the central limit theorem is predicated on an assumption that internal sensory events follow an independent and identical distribution, which is not considered to be satisfied in a net sense in psychophysical measurements (Green & Swets, 1966, p. 58). Therefore this article did not venture to question the a priori plausibility of the frameworks, which seems not to lend itself to quantitative evaluation.

[2]It is the kurtosis parameter that matters in the present findings (see Supplementary Figures S2-S4 in Supplementary Material 2). Thus it is rather important to systematically understand the continuous contributions of kurtosis through the illustrations of these two influential frameworks. This is important as the internal distributional form can change from situation to situation (Green & Swets, 1966, p. 58), and it may not exactly be gaussian nor logistic in empirical experiments.

[3]For example, type 1 ROCs from Yes/No experiments typically exhibit asymmetric shapes, which invite difficulties in meta-SDT model fittings (see Maniscalco & Lau, 2014).

[4]The reported results were replicated by regression with raw estimated criteria (see Supplementary Material 2).

References

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

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

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3*, 186–205.

Denison, R. N., Adler, W. T., Carrasco, M., & Ma, W. J. (2018). Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 11090–11095.

Egan, J. P., & Clarke, F. R. (1956). Source and receiver behavior in the use of a criterion. *Journal of the Acoustical Society of America, 28*, 1267–1269.

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

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

Falmagne, J. C. (1985). *Elements of psychophysical theory*. New York, NY: Oxford University Press.

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,* 843–876.

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

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife, 5*, e13388.

Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: Testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1795–1804.

Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. *Psychological Review, 122*, 542–557.

Kellen, D., & Klauer, K. C. (2018). Elementary signal detection and threshold theory. In E.J.

Wagenmakers (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (4th ed., Vol. *V*). New York, NY: Wiley.

Kellen, D., Winiger, S., Dunn, J. C., & Singmann, H. (2021). Testing the foundations of signal detection theory in recognition memory. *Psychological Review, 128*, 1022–1050.

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

Kornbrot, D. E. (2006). Signal detection theory, the approach of choice: Model-based and distribution-free measures and evaluation. *Perception & Psychophysics, 68*, 393–414.

Kristensen, S. B., Sandberg, K., & Bibby, B. M. (2020). Regression methods for metacognitive sensitivity. *Journal of Mathematical Psychology, 94*, 102297.

Lee, A. L. F., Ruby, E., Giles, N., & Lau, H. (2018). Cross-domain association in metacognitive efficiency depends on first-order task types. *Frontiers in Psychology, 9*, 2464.

Luce, R. D. (1959). *Individual choice behavior.* New York, NY: Wiley.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

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

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta-d′, response-specific meta-d′, and the unequal variance SDT model. In S. M. Fleming, & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 25–66). New York, NY: Springer.

Maniscalco, B., & Lau, H. (2015). Manipulation of working memory contents selectively impairs metacognitive sensitivity in a concurrent visual discrimination task. *Neuroscience of Consciousness, 2015*, niv002.

Miyamoto, K., Osada, T., Setsuie, R., Takeda, M., Tamura, K., Adachi, Y., . . . Miyashita, Y. (2017). Causal neural network of metamemory for retrospection in primates. *Science, 355*, 188–193.

Miyoshi, K., Kuwahara, A., & Kawaguchi, J. (2018). Comparing the confidence calculation rules for forced-choice recognition memory: A winner-takes-all rule wins. *Journal of Memory and Language, 102*, 142–154.

Miyoshi, K., & Lau, H. (2020). A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychological Review, 127*, 655–671.

Myung, J. I., & Pitt, M. A. (2018). Model comparison in psychology. In E.J. Wagenmakers (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (4th ed., Vol. *V*). New York, NY: Wiley.

Norwich, K. H. (1993). Information, sensation, and perception. San Diego, CA: Academic Press.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review, 26*, 1596–1618.

Peters, M. A. K. (2021, April 27). Towards characterizing the canonical computations generating phenomenal experience. *PsyArXiv*, https://doi.org/10.31234/osf.io/bqfr6.

Pleskac, T. J. (2015). Decision and choice: Luce's choice axiom. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., pp. 895–900). Oxford: Elsevier.

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

Rahnev, D. (2021). A robust confidence-accuracy dissociation via criterion attraction. *Neuroscience of Consciousness, 2021*, niab039.

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., & Akdoğan, B., …Bègue, I. (2020). The confidence database. *Nature Human Behaviour, 4*, 317–325.

Rahnev, D., & Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness, 2019*, niz009.

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*, 697–719.

Rollwage, M., & Fleming, S. M. (2021). Confirmation bias is adaptive when coupled with efficient metacognition. *Philosophical Transactions of the Royal Society B, 376*, 20200131.

Rouder, J. N., Pratte, M. S., & Morey, R. D. (2010). Latent mnemonic strengths are latent: A comment on Mickes, Wixted, and Wais (2007). *Psychonomic Bulletin & Review, 17*, 427–435.

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*, 165–175.

Rouy, M., Saliou, P., Nalborczyk, L., Pereira, M., Roux, P., & Faivre, N. (2021). Systematic review and meta-analysis of metacognitive abilities in individuals with schizophrenia spectrum disorders.

Neuroscience & Biobehavioral Reviews, 126, 329–337.

Samaha, J., Iemi, L., & Postle, B. R. (2017). Prestimulus alpha-band power biases visual discrimination confidence, but not accuracy. *Consciousness and Cognition, 54*, 47–55.

Shekhar, M., & Rahnev, D. (2020). Sources of metacognitive inefficiency. *Trends in Cognitive Sciences, 25*, 1–12.

Shekhar, M., & Rahnev, D. (2021). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review, 128*, 45–70.

Spence, M. L., Mattingley, J. B., & Dux, P. E. (2018). Uncertainty information that is irrelevant for report impacts confidence judgments. *Journal of Experimental Psychology: Human Perception and Performance, 44*, 1981–1994.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*, 702–712.

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1955). *The evidence for a decision-making theory of visual detection*. Technical Report No. 40, Electronic Defense Group, University of Michigan.

Toscani, M., Mamassian, P., & Valsecchi, M. (2021). Underconfidence in peripheral vision. *Journal of Vision, 21*, 1–14.

Webb, T. W., Miyoshi, K., So, T. Y., Rajananda, S., & Lau, H. (2021, November 12). Performance-optimized neural networks as an explanatory framework for decision confidence. *bioRxiv*, https://doi.org/10.1101/2021.09.28.462081.

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*, 10–65.

Zylberberg, A., Roelfsema, P. R., & Sigman, M. (2014). Variance misperception explains illusions of confidence in simple perceptual decisions. *Consciousness and Cognition, 27*, 246–253.