Measures of metacognition on signal-detection theoretic models

Article (Accepted Version)

Measures of metacognition on signal-detection theoretic models

Adam B. Barrett, Zoltan Dienes and Anil K. Seth

University of Sussex

Running head: MEASURES OF METACOGNITION

Author note

Adam B. Barrett, Sackler Centre for Consciousness Science and Department of Informatics, University of Sussex, Brighton, UK; Zoltan Dienes, Sackler Centre for Consciousness Science and School of Psychology, University of Sussex, Brighton, UK; Anil K. Seth, Sackler Centre for Consciousness Science and Department of Informatics, University of Sussex, Brighton, UK.

This research was supported by a donation from the Dr. Mortimer and Theresa Sackler Foundation via the Sackler Centre for Consciousness Science, and by EPSRC Leadership Fellowship EP/G007543/1 to AKS. We are grateful to Drs. Daniel Bor and Ryan Scott for helpful comments.

Correspondence concerning this article should be addressed to Adam B. Barrett, Department of Informatics, University of Sussex, Brighton BN1 9QJ, UK. E-mail: adam.barrett@sussex.ac.uk

Abstract

Analysing metacognition, specifically knowledge of accuracy of internal perceptual, memorial or other knowledge states, is vital for many strands of psychology, including determining the accuracy of feelings of knowing, and discriminating conscious from unconscious cognition. Quantifying metacognitive sensitivity is however more challenging than quantifying basic stimulus sensitivity. Under popular signal detection theory (SDT) models for stimulus classification tasks, approaches based on type II receiver-operator characteristic (ROC) curves or type II *d*-prime risk confounding metacognition with response biases in either the type I (classification) or type II (metacognitive) tasks. A new approach introduces meta-d': the type I *d*-prime that would have led to the observed type II data had the subject used all the type I information. Here we (i) further establish the inconsistency of the type II *d*-prime and ROC approaches with new explicit analyses of the standard SDT model, and (ii) analyse, for the first time, the behaviour of meta-d' under non-trivial scenarios, such as when metacognitive judgments utilize enhanced or degraded versions of the type I evidence. Analytically, meta-d' values typically reflect the underlying model well, and are stable under changes in decision criteria; however, in relatively extreme cases meta-d' can become unstable. We explore bias and variance of in-sample measurements of meta-d' and supply MATLAB code for estimation in general cases. Our results support meta-d' as a useful measure of metacognition, and provide rigorous methodology for its application. Our recommendations are useful for any researchers interested in assessing metacognitive accuracy.

Keywords: metacognition; signal-detection theory; modeling; meta-d'; confidence; discrimination

Metacognition, and in particular the ability to assess the accuracy of knowledge states, is fundamental to understanding executive processes (e.g. Koriat, 2007), the nature of memory (e.g. Mazzoni, Scoboria, & Harvey, 2010), good educational practice (e.g. Koriat, 2012), gambling (e.g. Lueddeke & Higham, 2011), development (e.g. Beck, McColgan, Robinson & Rowley, 2011), cognitive differences between species (e.g. Smith, Beran, Couchman, Coutinho & Boomer, 2009), social interaction (e.g. Frith, 2012), mental illness (e.g. Hamm et al 2012), and the distinction between conscious and unconscious processes in perception (e.g. Kanai, Walsh, & Tseng, 2010) and learning (e.g. Dienes & Seth, 2010). Given the range of applications, it would be helpful to have standard guidelines on measures of metacognitive accuracy. Nelson (1984) argued for Goodman-Kruskal's gamma coefficient, G, as an all purpose measure of association in metacognition research. His arguments persuaded many researchers; Masson & Rotello (2009) reported that in 2000-2008, of 64 articles on metacognition in the journal titles they chose, half had followed Nelson's advice. Nonetheless, Masson and Rotello argued that G is sensitive to bias (i.e. *a priori* disposition to respond in one way or another), and thus not ideal. They recommended a signal detection approach instead. Here we will pursue the use of signal detection theory (SDT) to determine the suitability of the different signal detection measures of association available for assessing metacognition.

SDT has been a major innovation in psychology and neuroscience, proving extremely useful for measuring stimulus discrimination accuracy independently of response bias (Lau & Passingham, 2006; Lau, 2008; Macmillan & Creelman, 2005). In a typical stimulus discrimination study subjects encounter many trials, in each of which they make a forced-choice response, classifying a stimulus as either present versus absent, or as 'type A' versus 'type B'. SDT posits that the discrimination decision on this so-called 'type I' task is based on internally generated evidence that follows distinct Gaussian probability distributions in the respective scenarios of 'absent' and 'present' (see

Figure 1 and e.g. Macmillan & Creelman, 2005). The fundamental SDT measure of discrimination performance, 'type I *d*-prime', is defined theoretically as the difference between the means divided by the standard deviation of the 'absent' distribution.[1] SDT further assumes a decision threshold (or criterion) determining whether the subject responds 'absent' or 'present', allowing each trial to be classified as a *hit*, *miss*, *false alarm*, or *correct rejection*. On this model, type I *d*-prime is by definition independent of this decision threshold and is therefore insensitive to response bias.

Given the success of SDT in measuring type I stimulus discrimination, there has been a natural motivation to apply it also to the so-called type II, or metacognitive, task (Clarke, Birdsall, & W. P. Tanner, 1959; Galvin, Podd, Drga, & Whitmore, 2003; Maniscalco & Lau, 2012; Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010; Ko & Lau, 2012) in which the subject classifies their (type I) stimulus classification responses as either 'confident' or 'unconfident' (or as occupying a point on a continuous or discrete confidence scale), reflecting the extent to which s/he can discriminate between their correct and incorrect responses. On the standard SDT model, this confidence judgment (type II response) is made based on two confidence thresholds applied to the same evidence as for the type I response (see Figure 1, and Macmillan & Creelman, 2005; Kunimoto, Miller & Pashler, 2001). Within this general framework, several methods for measuring metacognition on such tasks have been proposed, however obtaining a measure that is stable and insensitive to (types I and II) response bias has proven challenging (Galvin et al., 2003, Masson & Rotello, 2009). In this paper we explore and evaluate three distinct SDT-inspired approaches to measuring metacognitive sensitivity, and present evidence favouring one in particular - the relatively new meta-d' measure (Maniscalco & Lau, 2012).

We first consider 'type II *d*-prime', which is computed from type II hit and false alarm rates using the directly analogous formula to type I *d*-prime (a type II hit [false alarm] is a correct [incorrect] type I response made with high confidence; Kunimoto et al., 2001). While superficially appealing, this approach has been criticised because the standard SDT model explicitly implies non-Gaussian distributions for correct and incorrect type I responses, violating the assumptions underlying the *d*-prime measure (Galvin et al., 2003; Evans & Azzopardi, 2007; Maniscalco & Lau, 2012). Nevertheless, a thorough theoretical analysis of the behaviour of this measure on this model has not been previously performed. Leveraging new formulae for type II quantities, we address this need by systematically exploring the sensitivity of type II *d*-prime to (types I and II) response bias by varying decision and confidence thresholds across their entire ranges.

An alternative to type II *d*-prime that has received some support (Kolb & Braun, 1995; Wilimzig, Tsuchiya, Fahle, Einhuser, & Koch, 2008; Clifford, Arabzadeh & Harris, 2008; Masson & Rotello, 2009) is the use of receiver operating characteristic (ROC) curves to assess type II behaviour. A type I ROC curve characterizes stimulus discriminability by plotting the hit rate against false alarm rate for all possible decision thresholds (see Figure 4 for examples). While type I ROC curves are easy to obtain and interpret, the type II case is less clear. Specifically, when plotting type II hit rate against type II false alarm rate, multiple type II ROC curves can be constructed because these curves depend on three parameters: the single type I threshold and the two type II thresholds. Galvin et al. (2003) and Clifford et al. (2008) have each proposed distinct approaches to constructing ROC curves for analysing metacognition. However both approaches yield results that depend on type I response bias. Here we undertake new systematic analyses involving derivations of ROC curves that show, for a given decision threshold, the *maximum attainable* type II hit rate as a function of type II false alarm rate.

The approach we consider in the most detail is the meta-d' approach, recently introduced by Maniscalco and Lau (Maniscalco & Lau, 2012; Rounis et al., 2010). The conceptual basis of meta-d' is to compute the type I *d*-prime that would have led to the observed type II data, had the observer been using the standard SDT model. Departure from optimal metacognition on the standard SDT model can then be measured by comparing meta-d' with type I d-prime. By construction meta-d' and type I d-prime are equal, irrespective of (types I and II) response bias, on the standard SDT model. However, its behaviour under more general and empirically plausible scenarios has not been previously explored. Addressing this need, we systematically analyse scenarios in which metacognitive judgments utilize enhanced or degraded versions of the type I evidence, and when the decision threshold is jittered (Kellen, Klauer, & Singmann, 2012). Further we perform simulations to characterize how the meta-d' measures behave with finite data samples.

Type I signal detection theory

FIGURE 1 HERE

In this section we introduce formally the standard feed-forward Gaussian SDT model of perception. This model is schematized in Figure 1, and a full list of mathematical notation and terminology is summarized in Table 1. On this model, the task is to correctly classify a binary stimulus $S \in \{0, 1\}$, with the value 0 corresponding to stimulus absent and the value 1 corresponding to stimulus present. (The case of instead distinguishing between two different stimuli is treated in Appendix A; all results are very similar.) Throughout we assume an equal likelihood of stimulus present or stimulus absent, so $P(S = 1) = 0.5$. Perception is based on the evidence X, which is a Gaussian random variable, with mean and standard deviation dependent on S. When the stimulus is absent X takes a standard Gaussian distribution of mean 0 and standard deviation 1, i.e. $\mathcal{N}(0, 1)$. We denote the cumulative distribution function of this probability distribution by Φ_0, and the probability density function by ϕ_0. Given the stimulus is present, the distribution for X is a Gaussian of mean d' and standard deviation σ. We denote the cumulative distribution function of this probability distribution by $\Phi_{d',\sigma}$, and the probability density function by $\phi_{d',\sigma}$ (see Figure 2). These conventions are consistent with a definition of d' as the distance between means in units of the noise distribution standard deviation.[2] The response R is also binary, with $R = 0$ corresponding to responding 'stimulus absent' and $R = 1$ corresponding to responding 'stimulus present'. The response is decided based upon the decision threshold, θ (also referred to as 'type I threshold'). If $X < \theta$ then $R = 0$, and if $X \geq \theta$ then $R = 1$.

TABLE 1 HERE

To measure d' from empirical data, we compute the hit rate $h =: P(R = 1|S = 1)$ (i.e. the probability that the response is 'present' given that the stimulus is present) and the false alarm rate $f =: P(R = 1|S = 0)$ (i.e. the probability that the response is 'present' given that the stimulus is absent), and utilize the formula

$$d' = \sigma\Phi_0^{-1}(h) - \Phi_0^{-1}(f)\,, \tag{1}$$

which follows from the easy-to-derive formulae:

$$h = 1 - \Phi_{d',\sigma}(\theta)\,, \tag{2}$$

$$f = 1 - \Phi_0(\theta)\,. \tag{3}$$

One can also obtain the decision threshold θ in terms of the false alarm rate:

$$\theta = -\Phi_0^{-1}(f)\,. \tag{4}$$

Note that the standard deviation σ of the 'stimulus present' distribution is an extra parameter that cannot be determined by a single measurement of the hit and false alarm rates. In practice $\sigma = 1$ is often assumed.[3] Alternatively, one can ask subjects to include a confidence rating with their response,[4] and then obtain by proxy multiple measurements of the hit and false alarm rates associated with alternative decision thresholds in which, for example, $R = 1$ is assigned only to trials on which the subject replies 'stimulus present' and gives a confidence rating above a certain value. Each corresponding hit and false alarm rate can then be substituted into (1), allowing the d' and σ that provide the best overall fit to be found. Throughout this paper, however, we assume σ is a fixed, given parameter, and hence that a single measurement of the hit and false alarm rates will determine d' and the decision threshold θ.

It is useful to note that, while d' is analytically invariant with respect to the decision threshold, in practice measurements of d' become numerically unstable for extreme decision thresholds. This is because if either the hit rate or the false alarm rate is close to either 0 or 1, a small inaccuracy in measurement leads to a drastic error in the corresponding Φ_0^{-1} value in the formula (1) (Figure 2).

FIGURE 2 HERE

Type II signal detection theory

The standard model of type II SDT is applicable to the type II task of correctly distinguishing correct from incorrect responses made under the type I SDT framework. Here, we provide expressions for type II hit and false alarm rates for use in subsequent sections. For a summary of mathematical notation and terminology see Table 1.

On the standard SDT model, the confidence judgment (type II response) is based on the same evidence as the type I response (Figure 1). We denote correctness of the type I response by T (for truth), and the confidence in the response by C. As for the stimulus and response, these variables are assumed to take binary values belonging to $\{0, 1\}$, with 1 corresponding respectively to the response being correct and the subject being confident in their response. Confidence thresholds (also referred to as type II thresholds) τ_- and τ_+ are introduced. If the evidence X is less than τ_- or greater than τ_+ then the subject is confident, $C = 1$. We constrain the types I and II thresholds to be in a sensible order, so $\tau_- < \theta < \tau_+$. The type II hit rate H and false alarm rate F are defined respectively as

$H =: P(C = 1|T = 1)$, $F =: P(C = 1|T = 0)$, i.e. rates for being confident when correct and for being confident when incorrect. Given these conventions the following formulae for H and F can be derived (see Appendix B for details):

$$H = \frac{1 + \Phi_0(\tau_-) - \Phi_{d',\sigma}(\tau_+)}{1 + \Phi_0(\theta) - \Phi_{d',\sigma}(\theta)}\,, \tag{5}$$

$$F = \frac{1 - \Phi_0(\tau_+) + \Phi_{d',\sigma}(\tau_-)}{1 - \Phi_0(\theta) + \Phi_{d',\sigma}(\theta)}\,. \tag{6}$$

Response conditional type II hit rates H_+ and H_- are defined as the type II hit rates restricted respectively to positive and negative (type I) responses,
i.e. $H_+ =: P(C = 1|T = 1, R = 1)$, $H_- =: P(C = 1|T = 1, R = 0)$. Similarly, response conditional type II false alarm rates F_+ and F_- are defined by
$F_+ =: P(C = 1|T = 0, R = 1)$, $F_- =: P(C = 1|T = 0, R = 0)$. The formulae for these quantities are:

$$H_+ = P(X > \tau_+|X > \theta, S = 1) = \frac{1 - \Phi_{d',\sigma}(\tau_+)}{1 - \Phi_{d',\sigma}(\theta)}\,, \tag{7}$$

$$F_+ = P(X > \tau_+|X > \theta, S = 0) = \frac{1 - \Phi_0(\tau_+)}{1 - \Phi_0(\theta)}\,, \tag{8}$$

$$H_- = \frac{\Phi_0(\tau_-)}{\Phi_0(\theta)}\,, \tag{9}$$

$$F_- = \frac{\Phi_{d',\sigma}(\tau_-)}{\Phi_{d',\sigma}(\theta)}\,. \tag{10}$$

In what follows, we will describe, analyse and evaluate three possible ways of measuring performance at the type II task: these are respectively type II d-prime, type II ROC curve analysis, and meta-d'. A good measure should not depend on decision or confidence thresholds, but only on the amount of information available for making the confidence judgment. Thus, when assuming that the type I and II responses are based on the same standard SDT model, a good measure should be fully determined by d'. In other scenarios (see for example the section 'Meta-d' on alternative SDT models'), a good measure should generally increase with the amount of information available for making the confidence judgment.

Type II d-prime

FIGURE 3 HERE

Type II *d*-prime, which we denote by D', is computed by simply substituting the type II hit and false alarm rates into the formula (1) for type I *d*-prime. As mentioned in the introduction, type II *d*-prime is not a principled measure of type II performance, since under standard SDT assumptions the type II decision axis cannot be mapped onto a variable for which the evidence is given by a pair of Gaussian distributions (i.e., of the form shown in Figure 1 with a transformed version of the type I evidence on the horizontal axis, see Galvin et al., 2003). Further, Evans & Azzopardi (2007) found type II *d*-prime to vary strongly with (types I and II) response bias in various empirical scenarios. Nevertheless, a systematic investigation into the theoretical behaviour of this measure on the standard SDT model, under varying decision and confidence thresholds, has not previously been performed. We provide this here.

From (5) and (6) type II *d*-prime on the standard SDT model is given in terms of decision and confidence thresholds by

$$D' \equiv \sigma\Phi_0^{-1}(H) - \Phi_0^{-1}(F) = \sigma\Phi_0^{-1}\left(\frac{1+\Phi_0(\tau_-)-\Phi_{d',\sigma}(\tau_+)}{1+\Phi_0(\theta)-\Phi_{d',\sigma}(\theta)}\right) - \Phi_0^{-1}\left(\frac{1-\Phi_0(\tau_+)+\Phi_{d',\sigma}(\tau_-)}{1-\Phi_0(\theta)+\Phi_{d',\sigma}(\theta)}\right). \tag{11}$$

Using this formula, D' can be computed for any values of d', σ, θ, τ_+ and τ_-. We investigated D' across the full space of possible type I and II thresholds for the case $d' = 1$, $\sigma = 1$. Figure 3 shows the behaviour of D' in an informative subset of these cases. To better understand how variations in D' arise in these scenarios, Figure 3 also shows the corresponding type II ROC curves (i.e., the relation between type II hit and false alarm rate under each of the variations of thresholds; bottom row). To enable later comparison of measures we also show meta-d', which is defined in the section 'Meta-d''.

Figure 3(d) shows that when $\theta = 0.5$ (i.e. at the point of intersection of the two evidence distributions) the highest D' is obtained by placing the type II thresholds τ_+ and τ_- as far away as possible from the type I threshold θ. That is, D' is maximized by being *maximally unconfident*. Figure 3(e) shows that it is possible to have D' negative when all of the type I and II thresholds are set high; while this scenario clearly does not reflect a sensible choice of decision and confidence criteria it is nevertheless not a priori obvious that D' could be negative. Again challenging intuition, Figure 3(f) shows that is possible to have D' greater than d'. Previously, this outcome had been shown to be impossible given the assumption that $F = f$ (Galvin et al., 2003); our results show that this result does not generalize to the more general case of $F \neq f$.[5]

Some of the more extreme D' values obtained in the above analyses arise from decision and confidence threshold values that lead to extreme (i.e., close to 0 or 1) hit rates or false alarm rates. However, D' remains very sensitive to decision and confidence threshold values even for empirically reasonable ranges for which $0.05 < h, f, H, F, H_+, F_+, H_-, F_- < 0.95$. For example, within these ranges one can obtain $D' = 0.40$ by taking $\tau_- = 0.4, \theta = 0.5, \tau_+ = 0.6$, and a very different $D' = 0.95$ by taking $\tau_- = -0.5, \theta = 1.6, \tau_+ = 1.65$. Reflecting these variations in D', the positions of the corresponding ROC curves (in relation to the type I ROC curve) are also highly variable [Figure 3(g)-(i)]. Type II ROC curves are examined in more detail in the section 'Optimal type II ROC curves and H_{max}'.

Summarizing results in this section, we have confirmed the hypothesis that D' is a poor measure of metacognitive sensitivity, validating previous empirical findings (Evans & Azzopardi, 2007). For the standard SDT model, D' is highly dependent on choice of decision and confidence thresholds, takes high values when confidence thresholds are such that confidence is almost always low [Figure 3(a,d)], and is not bounded from either above or below by the benchmark values of 0 [Figure 3(b,e)] or type I d-prime [Figure 3(c,f)].

Optimal type II ROC curves and $H_{\max}$

We next investigate the use of ROC curves to measure type II performance on the standard SDT model. ROC curves have been proposed as being useful for evaluating type II sensitivity in a more stable manner than D', since they characterize type II behaviour over a range of confidence thresholds (Galvin et al., 2003; Macmillan & Creelman, 2005). However, in contrast to type I SDT, for which there is a single decision threshold, and hence a single ROC curve, for type II SDT there are in addition two confidence thresholds that can be varied, implying corresponding families of ROC curves.

In their important paper, Galvin et al. (2003) obtained a single type II ROC curve for each possible type I threshold by employing a slightly different model in which a single confidence threshold based on the likelihood ratio of being correct versus incorrect on the type I task. Using this approach they found that type II ROC curves depend strongly on the type I threshold, with considerable variation in the area underneath the type II ROC curve. This result implies that type II performance measured via ROC curves is also strongly dependent on type I response bias, even under 'perfect' metacognition that is utilising all of the available type I information (Maniscalco & Lau, 2012). Clifford et al. (2008) proposed assessing metacognitive sensitivity by comparing the standard type I ROC curve with alternative type I ROC curves derived from confidence ratings. The latter are derived from various regroupings of the data, for example, by classifying responses as 'present' only when the subject responds 'present' with high confidence. Under perfect metacognition, all ROC curves coincide; in other cases they diverge. Again though, the degree of divergence is not in general independent of type I response bias.

In this section we report new analyses exploring the *optimal* type II ROC curves attainable for each type I threshold. That is, for a fixed type I threshold θ, the ROC curve we consider is that which plots F against the maximum possible H given F and θ. In order to study these optimal type II ROC curves, we must define the quantity

$H_{\max} \equiv H_{\max}(f, h, F)$ as the maximal H for a given F, f and h on the standard SDT model. An algorithm for computing $H_{\max}$ is described in Appendix C (and Appendix D describes the dependence of $H_{\max}$ on type I response bias). For a given d', σ and θ, fixed values of h and f are computed using (2) and (3). The optimal type II ROC curve is then obtained by computing $H_{\max}$ for these fixed values of f and h across varying values of F.

FIGURE 4 HERE

Following this approach, Figure 4 shows the dependence of the optimal type II ROC curve on the type I threshold, for the case $d' = 1$, $\sigma = 1$. Figure 4(c) presents the optimal type II ROC curve when the type I threshold is placed at the intersection point of the two evidence distributions ($\theta = 0.5$). There it lies below the type I ROC curve. Figure 4(d) meanwhile plots this for a very conservative type I detection criterion ($\theta = 3$). There the type I and optimal type II ROC curves are approximately equal, but with the optimal type II ROC curve lying partly above the corresponding type I ROC curve [see D' in panel (f)]. In the bottom panels in Figure 4, D' values corresponding to $H_{\max}$ are plotted against F, and as expected show considerable variability, even though in all cases performance is optimal. To enable comparison we also plot meta-d', which is defined in the section 'Meta-d''.

The above results cast doubt on the utility of ROC curves for characterizing metacognitive sensitivity by showing that, even under the standard SDT model, these curves exhibit strong dependence on type I response bias. Although only extreme choices of decision and confidence thresholds lead to type II ROC curve points lying above the type I ROC curve, we may also conclude that it is difficult to assess by straightforward inspection of type I and type II ROC curve profiles whether SDT is a good fit for a dataset, since the expected discrepancy between curves is different in different cases. Having said this, our novel method of computing optimal type II ROC curves via $H_{\max}$ does provide an algorithmic test for whether a particular dataset is adequately modelled by SDT. Specifically, one can compare the observed H with the value of $H_{\max}$ computed from the observed values of F, h and f: If $H > H_{\max}$ then the data are not plausibly fit by SDT.

Meta-d'

So far, we have explored the properties of type II d-prime and optimal type II ROC curves on the standard SDT model, concluding that the former is highly dependent on both type I and II response bias, and that the latter is dependent on type I response bias. A true measure of metacognitive sensitivity should not depend on response biases, i.e. be independent of both decision and confidence thresholds on the SDT model, assuming that type I and II responses are based on the same underlying evidence (Maniscalco & Lau, 2012). Addressing this need, Lau et al. (Rounis et al., 2010; Maniscalco & Lau, 2012) have recently introduced meta-d', a measure which is explicitly designed to be constant and equal to d' whenever the standard SDT model underlies both type I and II responses. Meta-d' is defined as the type I d-prime that would have led to the observed type II data, assuming the subject's response and confidence judgment both follow the standard SDT model. Thus, departure from ideal metacognition will correspond to a difference between meta-d' and d', the magnitude of which has a clear interpretation in units that correspond to the stimulus absent evidence standard deviation. (Type II d-prime is formulated in different units from type I d-prime, making it hard to directly compare, notwithstanding the issue of response bias sensitivity.) Specifically, imperfect metacognition will be indicated by meta-$d' < d'$; alternatively, enhanced metacognition (e.g., potentially reflecting accumulation of information between type I and type II responses) would be indicated by meta-$d' > d'$.

There are several possible operational definitions of meta-d'. In this paper we examine the performance of two versions: Lau et al's meta-d'-SSE ($\tilde{d}'_{\mathrm{SSE}}$; sum-square error), and the novel meta-d'-balance ($\tilde{d}'_{\mathrm{b}}$). (Here and in all discussion of meta-d', tildes denote all the 'meta' quantities that enter equations, i.e. model parameters and type I data that would have led to the observed type II data if the observer were a standard SDT observer.) Both measures rely on two pairs of equations, one pair obtained by considering type II performance following a positive type I response, and the other pair obtained by considering type II performance following a negative type I response. These equations can not in general be solved simultaneously. The data-driven $\tilde{d}'_{\mathrm{SSE}}$ measure is obtained by finding the closest fit, i.e. the value which minimizes the sum of the squares of the errors of all the equations. Our theory-driven $\tilde{d}'_{\mathrm{b}}$ measure operates from a slightly different approach, being defined as the weighted average of the solutions for the two pairs of equations, weighted according to the proportion of positive and negative type I responses.

We now formally define the two measures. The construction of both $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$ depend on the quantities $\tilde{d}'_{+}$ and $\tilde{d}'_{-}$, each of which in turn depend on the 'relative type I threshold', Θ. Let us define all these entities. The 'relative type I threshold' Θ is the ratio between θ and d'. Then, $\tilde{d}'_{+}$ is the type I d-prime that would have led to the observed type II data H_{+} and F_{+} under the Θ implied by the type I data h and f; analogously, $\tilde{d}'_{-}$ is the type I d-prime that would have led to the observed type II data H_{-} and F_{-} under the Θ implied by h and f. Lau et al. (Rounis et al., 2010; Maniscalco & Lau, 2012) derive $\tilde{d}'_{\mathrm{SSE}}$ by assuming $\tilde{d}'_{+} = \tilde{d}'_{-}$ and obtaining the common value that minimizes the sum of the squares of the errors in the combined system of equations for $\tilde{d}'_{+}$ and $\tilde{d}'_{-}$ (see below). By contrast, $\tilde{d}'_{\mathrm{b}}$ allows $\tilde{d}'_{+}$ and $\tilde{d}'_{-}$ to differ, enabling the derivation of an implicit form analytical expression for meta-d'. $\tilde{d}'_{\mathrm{b}}$ is then defined as the weighted mean of $\tilde{d}'_{+}$ and $\tilde{d}'_{-}$, weighted according to the respective proportion of positive and negative type I responses.

Both versions of meta-d' are formulated in terms of response-conditional type II hit and false alarm rates (H_{+}, F_{+}, H_{-} and F_{-}) rather than response-unconditional hit and false alarm rates (H and F) because there is no unique type I d' that yields a given H and

F under a given relative type I threshold. This is the same reason, in terms of degrees of freedom, why h, f and F do not uniquely determine H [see the section 'Optimal type II ROC curves and $H_{\max}$', and also (Maniscalco & Lau, 2012)]. Note that we do not report on the properties of $\tilde{d}'_+$ and $\tilde{d}'_-$ themselves because we found them to be unstable compared to $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$.

To derive an expression for $\tilde{d}'_{\mathrm{b}}$ we first need an expression for the relative type I threshold Θ:

$$\Theta =: \frac{\theta}{d'} = \frac{-\Phi_0^{-1}(f)}{\sigma\Phi_0^{-1}(h) - \Phi_0^{-1}(f)}\,, \tag{12}$$

which follows from (4) and (1). Then, for $\tilde{d}'_+$, the meta type I threshold satisfies

$$\tilde{\theta}_+ = \Theta\tilde{d}'_+\,, \tag{13}$$

where the subscript '+' indicates 'meta' quantities related to $\tilde{d}'_+$. Substituting this expression into (7) and (8) yields H_+ and F_+ in terms of meta quantities, furnishing the implicit equations

$$H_+ = \frac{1 - \Phi_{\tilde{d}'_+,\sigma}(\tilde{\tau}_{++})}{1 - \Phi_{\tilde{d}'_+,\sigma}(\Theta\tilde{d}'_+)}\,, \tag{14}$$

$$F_+ = \frac{1 - \Phi_0(\tilde{\tau}_{++})}{1 - \Phi_0(\Theta\tilde{d}'_+)}\,, \tag{15}$$

which uniquely specify $\tilde{d}'_+$, as well as $\tilde{\tau}_{++}$, given H_+, F_+, h, f and σ, and Θ via (12). One can also obtain the meta type I data from the analogues of (2) and (3), i.e. the type I hit rate and false alarm rate that would have led to the observed type II data on the ideal SDT model, using the observed relative type I threshold:

$$\tilde{h}_+ \;=\; 1 - \Phi_{\tilde{d}'_+,\sigma}(\tilde{\theta}_+)\,, \tag{16}$$

$$\tilde{f}_+ \;=\; 1 - \Phi_0(\tilde{\theta}_+)\,. \tag{17}$$

Similarly, the equations for $\tilde{d}'_-$ are

$$\begin{aligned}
\tilde{\theta}_- &= \Theta\tilde{d}'_- \,, && (18)\\
H_- &= \frac{\Phi_0(\tilde{\tau}_{--})}{\Phi_0(\Theta\tilde{d}'_-)} \,, && (19)\\
F_- &= \frac{\Phi_{\tilde{d}'_-,\sigma}(\tilde{\tau}_{--})}{\Phi_{\tilde{d}'_-,\sigma}(\Theta\tilde{d}'_-)} \,, && (20)
\end{aligned}$$

and the meta type I quantities $\tilde{h}_-$ and $\tilde{f}_-$ are given by

$$\begin{aligned}
\tilde{h}_- &= 1 - \Phi_{\tilde{d}'_-,\sigma}(\tilde{\theta}_-) \,, && (21)\\
\tilde{f}_- &= 1 - \Phi_0(\tilde{\theta}_-) \,. && (22)
\end{aligned}$$

Having obtained $\tilde{d}'_+$ and $\tilde{d}'_-$, $\tilde{d}'_{\mathrm{b}}$ is computed as the weighted average:

$$\tilde{d}'_{\mathrm{b}} = r\tilde{d}'_+ + (1-r)\tilde{d}'_- \,, \qquad (23)$$

where r is the probability of a positive type I response, given by

$$r = \frac{1}{2}(h+f) \,. \qquad (24)$$

As mentioned, Lau et al's $\tilde{d}'_{\mathrm{SSE}}$ measure takes a slightly different approach. To compute $\tilde{d}'_{\mathrm{SSE}}$ one adds error terms to the equations (14), (15), (19) and (20) for $\tilde{d}'_+$ and $\tilde{d}'_-$, and substitutes a common value, $\tilde{d}'$, for both $\tilde{d}'_+$ and $\tilde{d}'_-$:

$$\begin{aligned}
H_+ &= \frac{1-\Phi_{\tilde{d}',\sigma}(\tilde{\tau}_{++})}{1-\Phi_{\tilde{d}',\sigma}(\Theta\tilde{d}')} + \epsilon_1 \,, && (25)\\
F_+ &= \frac{1-\Phi_0(\tilde{\tau}_{++})}{1-\Phi_0(\Theta\tilde{d}')} + \epsilon_2 \,, && (26)\\
H_- &= \frac{\Phi_0(\tilde{\tau}_{--})}{\Phi_0(\Theta\tilde{d}')} + \epsilon_3 \,, && (27)\\
F_- &= \frac{\Phi_{\tilde{d}',\sigma}(\tilde{\tau}_{--})}{\Phi_{\tilde{d}',\sigma}(\Theta\tilde{d}')} + \epsilon_4 \,. && (28)
\end{aligned}$$

$\tilde{d}'_{\mathrm{SSE}}$ is then the value of $\tilde{d}'$ that minimizes the sum of the squares of the errors $\epsilon_1, \ldots, \epsilon_4$.

Basic properties of meta-d' support its use as a measure of metacognitive sensitivity. First, such a measure should clearly indicate 'perfect' metacognition, i.e., when all the available type I information is utilized in making a type II response. Reflecting this by design, meta-d' is always equal to d' when type I and II responses are made using the standard SDT model. This property is shown explicitly in Figures 3 and 4 under multiple scenarios of decision and confidence threshold variation; see also (Maniscalco & Lau, 2012).

Second, a useful measure should smoothly increase with type II hit rate and smoothly decrease with type II false alarm rate when other variables remain constant. This is indeed typically how both meta-d' measures behave: Figure 5(a) shows an example of meta-d' decreasing with $F_\pm$, and Figure 5(b) shows an example of meta-d' increasing with $H_\pm$. The smooth dependence of meta-d' on $H_\pm$ and $F_\pm$ does however break down when either H_+, H_-, F_+, F_- or any of the meta type I hit or false alarm rates $\tilde{h}_+$, $\tilde{f}_+$, $\tilde{h}_-$, $\tilde{f}_-$ take very large or small values. In those cases, a small change in the data leads to large changes in the meta-d' measures. This instability occurs due to the nature of the Gaussian cumulative density function Φ_0 close to 0 or 1, and occurs also in standard (type I) SDT (Figure 2). Measurements of meta-d' will therefore sometimes be unstable, so it is useful to define exclusion criteria that lead to a restricted domain of application. Since meta-d' depends on both type I and type II quantities, a principled criterion would be to restrict application of meta-d' to cases for which

$$0.05 < \tilde{h}_+, \tilde{f}_+, \tilde{h}_-, \tilde{f}_-, H_+, F_+, H_-, F_- < 0.95 \,. \tag{29}$$

Note that it is necessary to include the meta type I quantities in this criterion, and that a more simple exclusion band based solely on the type II hit and false alarm rates would not be sufficient to ensure stability of the meta-d' measures. An example of meta-d' becoming unstable due to extreme meta type I quantities is that, in the entire wide exclusion area on the right-hand side of Figure 5(b), the meta type I false alarm rate $\tilde{f}_+$ is less than 0.05, even though in most of this region the type II hit and false alarm rates take middling values. More generally instability will arise at very high levels of type II response accuracy, when $H_{\pm}$ is much greater than $F_{\pm}$. The examples in Figure 5 show that when (29) is satisfied, the meta-d' measures vary slowly and smoothly as the data change, and that $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$ are in good agreement with each other.

FIGURE 5 HERE

Meta-d' on alternative SDT models

Previously, the theoretical behaviour of meta-d' has only been analysed on the standard SDT model, which assumes type I and II responses are made based on the same evidence, entailing 'perfect metacognition' (Maniscalco & Lau, 2012). However, in experimental data, departure from this behaviour has been observed, most notably with meta-d' less than d', interpreted as signalling imperfect metacognition (Maniscalco & Lau, 2012; Rounis et al., 2010). To examine the extent to which meta-d' measures do indeed capture metacognitive efficacy in a meaningful way, independently of (type I or II) response bias, we investigated the analytic behaviour of meta-d' measures on several distinct models that depart from the standard model.

FIGURE 6 HERE

We first considered a 'degrading signal' model, on which the type II response is based on weaker evidence than the type I response. On this model the type I response arises from evidence based on a standard SDT model, while the type II response is based on a regression of this evidence towards the mean of the stimulus absent distribution (i.e. 0), reflecting a possible time delay between the type I and II response (see Supplemental Material for details). Thus, given the type I evidence, the mean type II evidence is closer to zero than the type I evidence, and also has additional variance reflecting stochasticity in trial-to-trial signal decay. Figure 6 illustrates an example of this model and plots the behaviour of $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$. Both $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$ are approximately independent of decision and confidence thresholds. There is some variation as the type I threshold θ is varied, but this variation is very much smaller than common cases of type II d-prime variation on the standard SDT model (compare Figures 3 and 4).

FIGURE 7 HERE

We next considered an 'enhancing signal' model, on which the type II response is based on stronger evidence than the type I response. Again, the type I response arises from evidence based on a standard SDT model, but here the type II response is based on the assumption that the evidence accumulates over time (for 'present' stimuli; see Supplemental Material for details). One scenario in which this model might apply is when a subject has very limited time to make the type I response, but greater time to make the type II response. Figure 7 illustrates an example and plots the behaviour of $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$. Although both measures show some variation as decision and confidence thresholds are varied, both measures produce the expected output of meta-$d' > d'$ in all cases, reflecting the enhanced type II evidence. Further, the two measures give very similar values, with only slight divergence near the limits of allowed threshold ranges.

Further examples of the degrading and enhancing signal models, with unequal variances ($\sigma = 2$), are presented respectively in Figures S1 and S2 in the Supplemental Material. In those examples, the measures exhibit greater variability, and there are cases for which the $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$ are slightly divergent. However, importantly, in almost all cases, meta-$d' < d'$ on the degrading signal model and meta-$d' > d'$ on the enhancing signal model, reflecting respectively the degraded and enhanced type II evidence. The Supplemental Material also presents analyses of a model with criterion jitter (Kellen et al., 2012).

Our analyses of several alternative SDT models have confirmed that, across a broad range of empirically plausible scenarios, the meta-d' measures $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$ behave well as measures of metacognitive sensitivity. We found meta-d' measures to consistently give values that reflect the level of metacognition incorporated by design into each model. Moreover, the values obtained remained approximately invariant under changes to the decision and confidence thresholds, confirming that the measures are indeed quantifying metacognitive sensitivity independently of (type I or II) response bias and type I criterion jitter.

Bias and variance of meta-d' measures in sample

The results so far have considered the analytical behaviour of meta-d' on simple, idealized signal detection theoretic models. While very useful, such idealized models are not able to illuminate some issues of empirical importance, notably the possibility of bias and high variance when estimating meta-d' in sample. In this section we confront the behaviour of meta-d' measures on finite data samples by examining bias and variance in measurements of d' and meta-d' in simulation. We illustrate selected examples, that demonstrate possible outcomes at three different levels of assumed metacognitive performance. We then describe the MATLAB code provided with this paper, which can be used to generate simulated data allowing estimations of expected bias and variance for discrimination experiments in general.

To ensure generality our simulations do not assume a specific model of evidence or decision axes, rather they take as input just the type I and type II hit and false alarm rates h, f, $H_\pm$ and $F_\pm$. Given these values, we simulate a finite number of trials (50 per subject in the following examples), such that in-sample empirical hit and false alarm rates can be obtained. These empirical rates are then used to compute d' and meta-d' measures.

We have already demonstrated the importance of using exclusion criteria to ensure stability of SDT measures. These criteria become even more acute in finite samples. We therefore consider two alternative criteria for excluding outlying subjects in the following simulations. The first is a 'narrow' set which allow outliers, but which requires that hit and false alarm rates do not take their end values, i.e., we impose strict inequalities:

$$0 < \hat{h}, \hat{f}, \hat{H}_+, \hat{F}_+, \hat{H}_-, \hat{F}_- < 1, \quad \hat{h} \neq \hat{f}\,, \tag{30}$$

where ˆ denotes an empirical quantity. This is the minimal set of exclusion criteria for which meta-d' will be computable for all included subjects.

The second 'wide' set of exclusion criteria is based on (29), and excludes subjects with extreme data that can lead to distorted measurements due to the properties of z-values at limits (see Figure 2):

$$0.05 < \hat{h}, \hat{f}, \hat{\tilde{h}}_+, \hat{\tilde{f}}_+, \hat{\tilde{h}}_-, \hat{\tilde{f}}_- \hat{H}_+, \hat{F}_+, \hat{H}_-, \hat{F}_- < 0.95, \quad \hat{h} \neq \hat{f}\,. \tag{31}$$

TABLE 2 HERE

We explored bias and variance of d', $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$ in sample for low, medium and high metacognition examples. The chosen values for the type I and II hit and false alarm rates in each example are given in Table 2. For each example, and respectively for wide and narrow exclusion criteria, we simulated 50 trials per subject for 10,000 (non-excluded) subjects. For each level of metacognition and for each set of exclusion criteria we computed the probability of a subject being excluded, and the bias and standard deviation of measurements of d', $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$ across subjects.

FIGURE 8 HERE

Figure 8 shows the results from these simulations. Both meta-d' measures exhibited bias when estimated from finite data samples; in these simulations $\tilde{d}'_{\mathrm{SSE}}$ showed less bias than $\tilde{d}'_{\mathrm{b}}$. The bias was positive for narrow exclusion criteria and negative for wide exclusion criteria, with absolute values larger for the wide exclusion criteria. Type I d' was unbiased when using narrow exclusion criteria but became biased for wide exclusion criteria. The increase in bias for wider exclusion criteria can be explained by the exclusions being applied asymmetrically to outliers.

Both meta-d' measures exhibited greater standard deviation across subjects than d'. For wide exclusion criteria, $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$ showed similar variance, but for narrow exclusion criteria $\tilde{d}'_{\mathrm{b}}$ had a higher variance. Out of the three measures, only $\tilde{d}'_{\mathrm{b}}$ showed substantially less variance for the wider exclusion criteria compared with the narrow exclusion criteria. The decrease in variance for $\tilde{d}'_{\mathrm{b}}$ can be explained by the exclusions successfully removing extreme values.

Why are bias and variance different for the two different measures $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$? We suggest that the two measures can both be thought of as averages of two response-conditional meta-d' values, (one for positive type I responses and one for negative type I responses). The averaging is performed differently on each of the two measures (see section 'Meta-d''). In these examples, the effective weighting on $\tilde{d}'_{\mathrm{SSE}}$ is leading to lower bias and variance than $\tilde{d}'_{\mathrm{b}}$. However, it is unclear if this generalizes, particularly as Figure S1 shows an example for which $\tilde{d}'_{\mathrm{SSE}}$ is in theory less stable than $\tilde{d}'_{\mathrm{b}}$.

An alternative way of dealing with outlying subjects is to add 0.5, i.e. a 'flattening constant', to every data cell for every subject (Snodgrass & Corwin, 1988). This can be justified from a Bayesian perspective as the implementation of a prior belief that d' and decision and confidence thresholds are near zero, worth one observation for each of the hit and false alarm rates and corresponding miss and correct rejection rates, i.e. it is a unit information prior (Kass & Wasserman, 1995), corresponding to the belief that with approximately 95% probability all of the hit and false alarm rates lie between 0.05 and 0.95.[6] Such a prior, though vague, adds some information and can increase the accuracy of estimates when data are limited (Agresti & Coull, 1998; Greenland, 2006). The effects of flattening constants on the bias and variance of meta-d' measurements can be determined for a given experimental situation using the free code we describe below.

In summary, the above simulations show that bias and variance should be taken into account when measuring meta-d' from finite data sets. Narrow exclusion criteria lead to less bias but greater variance than wide exclusion criteria. For the examples considered here, the narrow exclusion criteria appear to perform better, particularly as the wide criteria can lead to many more subjects being excluded [Figure 8(a)]. However, in another example scenario that we ran, we simulated a high number (300) of trials per subject, and found that the bias was roughly zero independent of exclusion criteria, but again that a much smaller variance could be obtained with wide exclusion criteria. For a given experimental scenario, we therefore recommend using our simulation code (see below) to decide on exclusion criteria and gain expectations about bias and variance. Finally we note that in a paradigm in which subjects give a confidence rating on a multi-point (i.e. greater than two-point) scale, multiple readings of the meta-d' measures can be obtained: as with type I d' analyses, multiple measurements of each hit and false alarm rate would be obtained by using multiple thresholds for defining high and low confidence from the multi-point confidence scale. In that scenario, the simulations described here could be used to obtain estimates of bias and error separately for each such meta-d' reading, paving the way for finding the best way of combining the readings into a single estimate of the true meta-d' value. We will explore this in future work.

Simulation code

For application beyond the present study, we provide MATLAB simulation code (see ‘Supplemental material’) which furnishes estimates of the expected bias and variance of meta-d' for the number of trials per subject from which the data is drawn, and for the hit and false alarm rates actually observed. (Since hit and false alarm rates are binomially distributed under the assumption of independent identical trials, estimates of these from data are unbiased.) Knowing the expected bias is useful for obtaining more accurate estimates of meta-d', while knowing the expected variance is useful as it could be compared with the observed variance to derive an estimate of the variance of true meta-d'

values across subjects in an experiment.

Discussion

In this paper we have examined three distinct approaches to measuring metacognitive sensitivity within the framework of SDT. Corroborating previous analyses (Galvin et al., 2003; Evans & Azzopardi, 2007), we found that the type II *d*-prime and ROC curve approaches both risk confounding metacognition with (types I and II) response bias, and as a consequence can give misleading results. By contrast, our detailed analyses of the meta-d' approach support its use in measuring metacognition independently of other processes. Our specific contributions include (i) rigorous analytical characterization of the limits of type II *d*-prime and ROC curve analysis; (ii) definition of a new ROC curve analysis based on optimal type II ROC curves, furnishing a useful method for assessing whether SDT is an appropriate framework for modelling empirical data; (iii) derivation of (implicit form) analytical expressions for a new version of meta-d', $\tilde{d}'_{\mathrm{b}}$; (iv) rigorous examination of the behaviour of meta-d' under both standard SDT models and in empirically plausible scenarios involving signal degradation, signal enhancement, and trial-by-trial type I criterion jitter; (v) characterization of bias and variance in estimation of meta-d' measures in sample, and (vi) provision of easy-to-use MATLAB code enabling bias and variance estimation in a wide range of experimental and modelling situations.

Type II d-prime and ROC curves

In new systematic analyses of its behaviour on the standard SDT model, we have shown that type II *d*-prime (D') is highly dependent on decision and confidence thresholds, and in extreme cases can be negative, or even greater than (type I) d'. These findings corroborate empirical analyses by Evans & Azzopardi (2007). We found that even for empirically reasonable ranges of decision and confidence thresholds, D' values can vary across a range greater than $\frac{1}{2}d'$. Moreover, D' is typically maximized by being maximally unconfident on each trial. The apparently counterintuitive behaviour of D' in these examples is a direct consequence of incorrectly assuming Gaussian distributions for the evidence underlying confidence judgments. Further we found more generally that the discrepancy between types I and II ROC curves depended quite strongly on the type I threshold, even for optimal type II ROC curves (constructed for a given type I threshold by optimizing the type II hit rate for each type II false alarm rate). Somewhat counterintuitively, the maximum area under the type II ROC curve is attained by using an extreme type I threshold such that the type I response is the same on almost all trials. We conclude that metacognition cannot be indexed directly by the area under the type II ROC curve. A discrepancy between the observed type II ROC curve and the optimal type II ROC curve doesn't distinguish imperfect metacognition from a strategy which (as compared to maximizing type II performance) involves maintaining confidence thresholds separately for each type I response. (For example, setting a low confidence threshold on trials in which the type I response is 'present', and a higher confidence threshold on trials in which the type I response is 'absent' may result in an overall type II hit rate that is not optimal given the overall type II false alarm rate.)

Meta-d'

Meta-d' measures are explicitly designed to be exactly equal to d' whenever type I and II responses are made based on a standard SDT model, as in Figure 1. Thus, trivially on the standard SDT model, meta-d' is fully independent of (type I or II) response bias. Furthermore, any observed difference between meta-d' and d' has a clear interpretation in units that correspond to the stimulus absent evidence standard deviation. Despite these attractive properties, the consistency, stability and independence from response bias of meta-d' had not been previously examined beyond the standard SDT model. Here, we have confirmed that meta-d' remains consistent, stable, and mostly independent from response bias in alternative SDT models involving (i) a degrading (type II) signal, (ii) an enhancing (type II) signal, and (iii) (type I) criterion jitter; in all cases subject to decision and confidence thresholds lying within an empirically reasonable range. These results strongly support the use of meta-d' as a sensitive measure of metacognitive performance in discrimination tasks.

It is important to recognize that the concept of meta-d' can be operationalized in a variety of ways. In their introduction of the concept, Lau et al. took a data-driven approach (Rounis et al., 2010, Maniscalco & Lau, 2012; see section 'Meta-d''). Assuming a common level of metacognition following positive and negative type I responses, a single estimate of meta-d' is derived from a system of equations based on all of the observed hit rates and false alarm rates. This system of equations does not generally have a simultaneous solution, and so the estimate is obtained by minimizing the sum of the squares of the errors of the equations. This approach defines the measure $\tilde{d}'_{\mathrm{SSE}}$. (Lau et al. have also introduced a second method for fitting the equations, namely via a maximum likelihood approach, yielding the measure meta-d'-MLE. We have not analyzed meta-d'-MLE here, but we believe it to behave similarly to the versions of meta-d' that we have analyzed.) In this paper we have formulated an alternative measure which does not assume that the level of metacognition is the same following a positive or negative type I response. This new measure, $\tilde{d}'_{\mathrm{b}}$, is built from the weighted average (23) of respective values computed from positive and negative type I responses. For a given theoretical scenario, e.g. in our studies of the degrading signal and enhanced signal models, $\tilde{d}'_{\mathrm{b}}$ always has a well-defined value, that can be computed without any error, as the unique solution to our system of equations (14, 15, 19, 20, 23). While we have introduced the new $\tilde{d}'_{\mathrm{b}}$ measure because we consider it conceptually more straightforward for theoretical work, in practice both $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$ are equally usable for theoretical and empirical studies. True theoretical values of hit and false alarm rates can be plugged into the equations for $\tilde{d}'_{\mathrm{SSE}}$

and the value that minimizes the sum-square-error can be computed. This can then be taken to be the true value of $\tilde{d}'_{\mathrm{SSE}}$ for the model, even though it has been derived by considering error terms in a set of equations. We have indeed explored the theoretical behaviour of both measures, and their behaviour on finite samples of simulated data; we found $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$ to behave similar to each other in the scenarios we have considered.

Still other varieties of meta-d' can be envisaged. For example, one could also consider the response-conditional quantities $\tilde{d}'_{+}$ and $\tilde{d}'_{-}$ as separate measures in their own right. Interestingly, when applied to a perceptual detection task these measures could dissect differences in metacognition separately for 'seen' and 'unseen' trials (c.f. Kanai et al., 2010). However, initial exploration of these measures indicated that they are less stable than either $\tilde{d}'_{\mathrm{b}}$ or $\tilde{d}'_{\mathrm{SSE}}$; further investigation is beyond the present scope. We also defer for future investigations other varieties of meta-d' that fix quantities other than the relative type I threshold (Θ) when comparing type I and type II responses: possibilities include fixing the type I false alarm rate (i.e., meta-d' as the d' that would have led to the observed type II data at the same type I false alarm rate as that observed).

Our study is also the first to examine the statistical properties of meta-d' measures on finite data via a simulation model. By doing so we were able to characterize expected bias and standard error of meta-d' for finite subjects and trials-per-subject, as a function of type I and type II hit and false alarm rates. We illustrated this process in three simulation scenarios, representing 'low', 'medium', and 'high' metacognition respectively. The results were complex but can be summarized as showing: (i) variance in meta-d' is in general higher than variance in type I d-prime; (ii) bias could be positive or negative depending on the exclusion criteria employed, with narrow exclusion criteria (i.e., excluding fewer subjects) showing less bias but more variance as compared to wide exclusion criteria. Since bias and variance depend on many factors it is sensible to estimate these quantities for any given experiment, and use the estimates to tailor analyses and inferences. Facilitating this, along with this paper we provide MATLAB simulation code which furnishes estimates of bias and variance in meta-d' when given as inputs: (i) number of trials per subject, and (ii) observed hit and false alarm rates. This code will facilitate interpretation of meta-d' in discrimination experiments and can also be used to inform experiment design to optimize (i.e., minimize or trade-off) expected bias and variance in meta-d'.

Summary and conclusions

Measuring metacognitive (type II) performance independently from discrimination (type I) performance is a key challenge in any situation where metacognitive accuracy is important to assess in cognitive, developmental, social, educational, comparative, or abnormal psychology (Macmillan & Creelman, 2005, Beran, Brandl, Perner & Proust, 2012, Efklides & Misailidi 2010). For example, in consciousness research, metacognition is often interpreted as reflecting conscious processing (Kolb & Braun, 1995; Lau & Passingham, 2006; Szczepanowski & Pessoa, 2007; Rounis et al., 2010; Dienes & Perner, 1999; Dienes & Scott, 2005; Persaud, McLeod, & Cowey, 2007). Commonly, at-chance type II performance accompanying above-chance type I performance is taken to reflect implicit processing (Weiskrantz, 1997). Conversely, above-chance type II performance (accompanying above-chance type I performance) is taken to reflect conscious processing. Going further, it is tempting to interpret any discrepancy between performance at the two levels (when both are above chance) as the degree to which explicit (metacognitive) awareness falls short of implicit (unconscious) performance (Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008). Validating and quantifying claims like these requires type II measures that are sensitive to type II performance but independent of response biases at both type I and type II levels, and which are commensurable with type I measures. In this paper we have extensively reviewed and analysed several current approaches to this challenge, all grounded in SDT. Our results indicate that the relatively new measure, meta-d', is both theoretically the most principled and empirically the most useful measure among those considered. Complementing this conclusion we provide new (implicit) analytic formulae for calculating a new version of meta-d' ($\tilde{d}'_{\mathrm{b}}$), and simulation code which can be used to estimate sample bias and variance in meta-d'. Taken together our results provide important new constraints and new heuristics governing the design and interpretation of experiments involving measurements of metacognitive performance.

Some theories of consciousness emphasize a central role for metacognition. Notably, so-called higher-order-thought (HOT) theories (Rosenthal, 2005; Carruthers, 1996; Gennaro, 2004) propose that conscious content is specified by the existence of higher-order (i.e., metacognitive) representations of the corresponding first-order content. On these theories, metacognition is constitutively determinate of consciousness and measures of metacognition therefore represent clear operationalizations of the corresponding theories. For example, an inability to discriminate states of completely guessing from states of having some knowledge is good evidence that one is not aware of one's knowledge, and hence, on higher order theories, that the knowledge is unconscious. On these theories, the type II criterion should be distinguishing complete guessing from any amount of confidence in order for a zero meta-d' to show unconscious knowledge; conversely, having a zero meta-d' when confidence is always high is not diagnostic of unconscious knowledge (Dienes, 2004). However, one need not buy into HOT-type theories in order to benefit from reliable measures of metacognition for advancing our understanding of consciousness. All theories of consciousness rely either explicitly or implicitly on subjective reports as data, and delineating the boundaries between what happens implicitly and what is (reportably) conscious provides important constraints on any theory (Seth et al., 2008). One useful avenue to integrating consciousness theories with SDT will be to identify the extent to which separable brain networks subserve type I and type II discriminations

(Fleming, Weil, Nagy, Dolan, & Rees, 2010; Fleming & Dolan, 2012).

An important future challenge lies in integrating SDT with the increasingly influential framework of 'predictive coding' or the so-called 'Bayesian Brain' hypothesis (Rao & Ballard, 1999; Bubic, Cramon, & Schubotz, 2010; Friston, 2010; Clark, in press). According to this framework, perceptual content is determined by top-down predictive signals arising from multi-level generative models of the external causes of sensory signals, which are continually modified by bottom-up prediction error signals communicating mismatches between predicted and actual signals across hierarchical levels. This view stands in contrast to classical 'evidence accumulation' frameworks, exemplified by SDT, which treat bottom-up signals as carrying content. Although in some situations predictive coding and evidence accumulation are mathematically equivalent (Spratling, 2008), when considered from the perspective of SDT there remain many possible ways in which top-down expectations may shape evidence distributions at type I and type II levels (Turner, van Zandt, & Brown, 2011; Wyart, Nobre, & Summerfield, 2012). We hope the rigorous treatment of current SDT provided here will provide a firm platform from which these more speculative issues can be usefully explored.

Supplemental material

The file 'simmetadb.m' is a MATLAB m-file for estimating the bias and variance of d' and meta-d' measurements, as described in the section 'Bias and variance of meta-d' measures in sample'. The file takes as inputs mean type I and II hit and false alarm rates, the ratio of standard deviations σ, number of trials (s) per subject and number of subjects (n). Users also have the choice of whether to use 'narrow' or 'wide' criteria for subject exclusion, and whether to use a flattening constant (see section 'Bias and variance of meta-d' measures in sample'). The code simulates subjects performing s trials, computing empirical d' and $\tilde{d}'_{\mathrm{b}}$ for each subject, repeating until n non-excluded subjects have been simulated. The code outputs true values of d' and $\tilde{d}'_{\mathrm{b}}$, mean empirical d' and $\tilde{d}'_{\mathrm{b}}$, empirical standard deviation of d' and $\tilde{d}'_{\mathrm{b}}$, and the proportion of excluded subjects. The file 'metadprimepm.m' computes the standard estimate of $\tilde{d}'_{\mathrm{b}}$ from single subject data, taking as inputs the type I and II hit and false alarm rates and the assumed value of σ. The files 'eqformetadplus.m' and 'eqformetadminus.m' are auxiliary files.

The file 'furtherdetailsandresults.pdf' contains further material on meta-d' on alternative SDT models, namely degrading signal, enhancing signal and criterion jitter models.

References

Agresti, A. & Coull, B. A. (1998). Approximate Is better than ”exact” for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119-126.

Beck, S. R., McColgan, K. L. T., Robinson, E. J., & Rowley, M. G. (2011). Imagining what might be: why children under-estimate uncertainty. *Journal of Experimental Child Psychology, 110*, 603-610. doi: 10.1016/j.jecp.2011.06.010

Beran, M., Brandl, J. L. , Perner, J. & Proust, J. (Eds), (2012). *The Foundations of Metacogntion.* Oxford, England: Oxford University Press

Bubic, A., Cramon, D. Y. V., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4*(00025). doi: 10.3389/fnhum.2010.00025.

Carruthers, P. (1996). *Language, thought and consciousness.* Cambridge, England: Cambridge University Press.

Clark, A. (in press). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *xx*, xxx.

Clarke, F. R., Birdsall, T. G., & Tanner, W. P., Jr. (1959). Two types of ROC curves and definitions of parameters. *The Journal of the Acoustical Society of America*, *31*(5), 629-630. doi: 10.1121/1.1907764

Clifford, C. W. G., Arabzadeh, E., & Harris, J. A. (2008). Getting technical about awareness. *Trends in Cognitive Sciences*, *12*(2), 54-58. doi: 10.1016/j.tics.2007.11.009

Dienes, Z. (2004). Assumptions of subjective measures of unconscious mental states: Higher order thoughts and bias. *Journal of Consciousness Studies*, *11*, 25-45.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioural and Brain Sciences*, *22*(5), 735-755.

Dienes, Z., & Scott, R. B. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research*, *69*, 338-351. doi: 10.1007/s00426-004-0208-3

Dienes, Z., & Seth, A. K. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition*, *19*(2), 674-681. doi: 10.1016/j.concog.2009.09.009

Efklides, A. & Misailidi, P. (Eds) (2010). *Trends and Prospects in Metacognition Research.* New York, NY: Springer.

Evans, S. & Azzopardi, P. (2007). Evaluation of a 'bias-free' measure of awareness. *Spatial Vision*, *20*(1-2), 61-77. doi: 10.1163/156856807779369742

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1338-1349. doi: 10.1098/rstb.2011.0417

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. doi: 10.1126/science.1191883

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*, 127-138. doi: 10.1038/nrn2787

Frith, C. D. (2012). The role of metacognition in social interaction. *Philosophical Transactions of the Royal Society: B, 367*, 2213-2223. doi: 10.1098/rstb.2012.0123

Galvin, S., Podd, J., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin and Review*, *10*, 843-876. doi: 10.3758/BF03196546

Gennaro, R. J. (Ed.) (2004). *Higher-order theories of consciousness: An anthology.* Amsterdam, Netherlands: John-Benjamins.

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology*, *35*, 765-775. doi: 10.1093/ije/dyi312

Hamm, J. A., Renard, S. B., Fogley, R. L., Leonhardt, B. L., Dimaggio, G., Buck, K. D.

and Lysaker, P. H. (2012), Metacognition and Social Cognition in Schizophrenia: Stability and Relationship to Concurrent and Prospective Symptom Assessments. *Journal of Clinical Psychology.* doi: 10.1002/jclp.21906

Kanai, R., Walsh, V., & Tseng, C. H. (2010). Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and Cognition*, *19*(4), 1045 - 1057. doi: 10.1016/j.concog.2010.06.003

Kass, R. E., & Wasserman, L. A. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwartz criterion. *Journal of the American Statistical Association*, *90*, 928-934.

Kellen, D., Klauer, K. C., & Singmann, H. (2012). On the measurement of criterion noise in signal detection theory: The case of recognition memory. *Psychological Review*, *119*, 457-479. doi: 10.1037/a0027727

Ko, Y., & Lau, H. (2012). A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1401-1411. doi: 10.1098/rstb.2011.0380

Kolb, F. C., & Braun, J. (1995). Blindsight in normal observers. *Nature*, *377*(6547), 336–338. doi: 10.1038/377336a0

Koriat, A. (2007). Remembering: Metacognitive monitoring and control processes. In H. L. Roediger, III, Y. Dudai, & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 243-246). New York: Oxford University press.

Koriat, A. (2012). The relationships between monitoring, regulation and performance. *Learning and Instruction*, *22*, 296-298. doi: 10.1016/j.learninstruc.2012.01.002

Kunimoto, C., Miller, J. G., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, *10*(3), 294–340. doi: 10.1006/ccog.2000.0494

Lau, & Passingham (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences USA*, *103*, 18763-18768. doi: 10.1073/pnas.0607716103

Lau (2008). A higher order Bayesian decision theory of consciousness. *Progress in Brain Research*, *168*, 35-48. doi: 10.1016/S0079-6123(07)68004-2

Lueddeke, S. and Higham, P. A. (2011) Expertise and gambling: Using type-2 signal detection theory to investigate differences between regular gamblers and non-gamblers. *Quarterly Journal of Experimental Psychology*, *64*(9), 1850-1871. doi: 10.1080/17470218.2011.584631

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Hove, England: Psychology Press.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422 - 430. doi: 10.1016/j.concog.2011.09.021

Masson, M. E. J. & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(2), 509-527. doi: 10.1037/a0014876

Mazzoni, G., Scoboria, A., & Harvey, L. (2010). Non-believed memories. *Psychological Science, 21*(9), 1334-1340. doi: 10.1177/0956797610379865

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109-133.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, *10*(2), 257-261. doi: 10.1038/nn1840

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79-87. doi: 10.1038/4580

Rosenthal, D. M. (2005). *Consciousness and mind.* Oxford, England: Clarendon.

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. doi: 10.1080/17588921003632529

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). *Trends in Cognitive Sciences*, *12*(8), 314-321. doi: 10.1016/j.tics.2008.04.008

Smith, J. D., Beran, M. J., Couchman, J. J., Coutinho, M. V. C., & Boomer, J. B. (2009). Animal metacognition: Problems and prospects. *Comparative Cognition and Behavior Reviews, 4*, 40-53.

Spratling, M. W. (2008). Reconciling predictive coding and biased competition models of cortical function. *Front Comput Neurosci*, *2*(4). doi: 10.3389/neuro.10.004.2008

Szczepanowski, R., & Pessoa, L. (2007). Fear perception: Can objective and subjective awareness measures be dissociated? *Journal of Vision*, *7*(4). doi: 10.1167/7.4.10

Turner, B. M., van Zandt, T., & Brown, S. (2011). A dynamic stimulus driven model of signal detection. *Psychological Review*, *118*(4), 583–613. doi: 10.1037/a0025191

Weiskrantz, L. (1997). *Consciousness lost and found: A neuropsychological exploration.* New York, NY: Oxford University Press.

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

Wyart, V., Nobre, A. C., & Summerfield, C. (in press). Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences USA*. doi: 10.1073/pnas.1120118109

Appendix A

Mathematical conventions

Here we justify our conventions and describe an alternative definition of d'. Suppose the true evidence for stimulus absent, Y_0, has a $\mathcal{N}(\mu_0, \sigma_0^2)$ distribution, and the true evidence for stimulus present, Y_1, has a $\mathcal{N}(\mu_1, \sigma_1^2)$ distribution, i.e., the most general Gaussian case. Then our convention is to define d' as

$$d' = \frac{\mu_1 - \mu_0}{\sigma_0}\,, \tag{32}$$

the difference between the means, in units of the noise distribution standard deviation. A simple linear transformation defines

$$X_0 =: \frac{Y_0 - \mu_0}{\sigma_0}\,, \qquad X_1 =: \frac{Y_1 - \mu_0}{\sigma_0}\,. \tag{33}$$

Then $X_0 \sim \mathcal{N}(0, 1)$ and $X_1 \sim \mathcal{N}(d', \sigma^2)$, where $\sigma = \sigma_1/\sigma_0$, recovering the conventions described in the main section.

Another definition of d' that has appeared in the literature (Macmillan & Creelman, 2005) is

$$d' =: \frac{\mu_1 - \mu_0}{\sqrt{\frac{1}{2}\left(\sigma_0^2 + \sigma_1^2\right)}}\,. \tag{34}$$

In this case, the transformation (33) leads to $X_0 \sim \mathcal{N}(0, 1)$ and $X_1 \sim \mathcal{N}(\mu, \sigma^2)$, where

$$\mu = \frac{\mu_1 - \mu_0}{\sigma_0}\,, \qquad \sigma = \frac{\sigma_1}{\sigma_0}\,, \tag{35}$$

and the formula for d' becomes

$$d' = \frac{\mu}{\sqrt{\frac{1}{2}\left(1 + \sigma^2\right)}}\,. \tag{36}$$

This formula, and the variables X_0 and X_1, with the parameters μ, σ define the most convenient formulation of the general SDT model for this alternative definition of d'.

The results presented in this paper are very similar for either of these two possible definitions of d', and in particular for the case $\sigma = 1$ the two definitions are exactly equivalent.

Appendix B

Computation of type II hit rate and false alarm rate

Here we explain how to derive the formulae (5) and (6) for the type II hit rate and false alarm rate on the standard SDT model. For the hit rate we have

$$\begin{aligned} H \equiv P(C=1|T=1) &= P(C=1|R=0,S=0)\cdot P(R=0,S=0|T=1) \\ &\quad + P(C=1|R=1,S=1)\cdot P(R=1,S=1|T=1)\,. \end{aligned} \tag{37}$$

The four quantities on the RHS of this expression can be written down as follows:

$$P(R=0,S=0|T=1) = \frac{P(R=0,S=0)}{P(R=0,S=0)+P(R=1,S=1)} \tag{38}$$

$$= \frac{\Phi_0(\theta)}{\Phi_0(\theta)+1-\Phi_{d',\sigma}(\theta)}\,, \tag{39}$$

$$P(C=1|R=0,S=0) = \frac{\Phi_0(\tau_-)}{\Phi_0(\theta)}\,, \tag{40}$$

$$P(R=1,S=1|T=1) = \frac{1-\Phi_{d',\sigma}(\theta)}{\Phi_0(\theta)+1-\Phi_{d',\sigma}(\theta)}\,, \tag{41}$$

$$P(C=1|R=1,S=1) = \frac{1-\Phi_{d',\sigma}(\tau_+)}{1-\Phi_{d',\sigma}(\theta)}\,. \tag{42}$$

The expression (5) follows by substitution of these four expressions into (37). The formula (6) is derived following the same method.

Appendix C

Computation of $H_{\max}$ on the standard SDT model

To compute $H_{\max}$, we have to maximize H with respect to τ_+ and τ_-, given the data F, f and h. It is convenient to rewrite the formulae (5) and (6) for H and F as

$$\begin{aligned} H &= \frac{1 + \Phi_0(\tau_-) - \Phi_{d',\sigma}(\tau_+)}{1 + h - f}\,, && (43)\\ F &= \frac{1 - \Phi_0(\tau_+) + \Phi_{d',\sigma}(\tau_-)}{1 - h + f}\,, && (44) \end{aligned}$$

where here we have used (2) and (3). Then maximizing H, whilst keeping F fixed means extremizing the following quantity, where λ is a Lagrange multiplier:

$$Y =: \frac{1 + \Phi_0(\tau_-) - \Phi_{d',\sigma}(\tau_+)}{1 + h - f} + \lambda \left[\frac{1 - \Phi_0(\tau_+) + \Phi_{d',\sigma}(\tau_-)}{1 - h + f} - F \right]. \qquad (45)$$

Setting partial derivatives of Y with respect to τ_+ and τ_- to zero yields the following equation:

$$\phi_0(\tau_+)\phi_0(\tau_-) = \phi_{d',\sigma}(\tau_+)\phi_{d',\sigma}(\tau_-)\,. \qquad (46)$$

For the usual case of $\sigma = 1$ this leads to

$$\begin{aligned} \tau_+ &= -\tau_- + d'\,, && (47)\\ F &= \frac{1}{1 - h + f}\left[1 - \Phi_0\left(-\tau_- + d'\right) + \Phi_{d',\sigma}(\tau_-)\right]\,, && (48)\\ H_{\max} &= \frac{1}{1 + h - f}\left[1 + \Phi_0(\tau_-) - \Phi_{d',\sigma}\left(-\tau_- + d'\right)\right]\,. && (49) \end{aligned}$$

We find τ_- by solving (48) numerically and then, using the obtained value, obtaining $H_{\max}$ from (49). When this yields results with decision and confidence thresholds not in the correct order $\tau_- < \theta < \tau_+$, then the optimum is at the boundary, i.e., if $\tau_+ < \theta$ then we reset $\tau_+ = \theta$ and compute τ_- and H directly from (43) and (44) to obtain

$$\begin{aligned} \tau_- &= \Phi^{-1}_{d',\sigma}[(1 - h + f)F - f]\,, && (50)\\ H_{\max} &= \frac{h + \Phi_0(\tau_-)}{1 + h - f}\,. && (51) \end{aligned}$$

Similarly for the case in which τ_- comes out as greater than θ, we reset $\tau_- = \theta$ and obtain

$$\tau_+ = \Phi_0^{-1}[2 - h - (1 - h + f)F]\,, \tag{52}$$

$$H_{\max} = \frac{2 - f - \Phi_{d',\sigma}(\tau_+)}{1 + h - f}\,. \tag{53}$$

For the general case of $\sigma \neq 1$ equation (46) is quadratic in τ_+ and τ_-, and often has no real solution. This leads to a more complicated analysis, with optimum values of τ_+ or τ_- often occurring at the boundary.

Appendix D

Optimal type II thresholds and variation of $H_{\max}$ with type I response bias

Here, we illustrate the dependence of optimal type II thresholds and $H_{\max}$ on type I response bias. Figure 9 shows optimal values of $\tau_{\pm}$ and $H_{\max}$ for the case $d' = 1$, $\sigma = 1$ and for the two values $F = 0.5$ and $F = 0.1$. Of note is that for values of θ in between the peaks of the 'present' and 'absent' evidence distributions, the optimal type II thresholds are symmetric about the point of intersection of the two distributions, and not symmetric about θ. Also worth noting is that when $H_{\max}$ is compared for different values of θ, $H_{\max}$ is minimized when θ is at its optimum position at the midpoint of the 'present' and 'absent' distributions, and $H_{\max}$ is maximized for values of θ that correspond to very strong type I response bias.

FIGURE 9 HERE

Footnotes

[1] An alternative definition is the difference between the means divided by the mean of the standard deviations of the 'absent' and 'present' evidence distributions.

[2] An alternative measure utilizes units of the mean standard deviation of the two distributions, and is in fact a more principled measure for the case of distinguishing between two stimuli as opposed to stimulus present versus stimulus absent. We discuss these issues in Appendix A.

[3] This assumption may be safer for an A vs B discrimination than for a present vs absent discrimination; see Appendix A.

[4] Note that confidence ratings were first introduced in SDT to obtain a better characterization of the type I model, with a distinct purpose to the current study on metacognition and the type II model.

[5] Galvin et al. (2003) showed that for $F = f$, and the likelihood ratio of stimuli monotonically increasing along the decision axis, the type I ROC curve is an upper bound for the type II ROC curve; the monotonically increasing likelihood ratio condition is satisfied for our ideal SDT model with $\sigma = 1$.

[6] This also corresponds to the prior belief with 95% confidence that d' and β lie roughly between ±3. There is no magic for using 0.5 as the number added to each cell; if for example one was 95% sure that d' lay between ±2, one could add 1 to each cell; or if 95% sure that d' lay between ±1.5, one could add 2.

Table 1

Table of mathematical notation and terminology

Symbol	Description / Terminology
S	stimulus
R	(type I) response
T	correctness of response
C	confidence judgement (i.e. type II response)
h, f	respectively type I hit and false alarm rate
H, F	respectively type II hit and false alarm rate
H_+, F_+	respectively type II hit and false alarm rates restricted to positive type I responses
H_-, F_-	respectively type II hit and false alarm rates restricted to negative type I responses
X	the evidence on which the response and confidence judgment are made
θ	decision (type I) threshold
τ_+	upper confidence (type II) threshold
τ_-	lower confidence (type II) threshold
σ	ratio of standard deviations of the evidence in the case $S = 1$ to the case $S = 0$
Φ_0, ϕ_0	respectively the cumulative distribution function and probability density functions of the standard Gaussian distribution of mean 0 and variance 1
$\Phi_{d,\sigma}$, $\phi_{d,\sigma}$	respectively the cumulative distribution function and probability density functions of the Gaussian distribution of mean d and variance σ
d'	type I d-prime
D'	type II d-prime
Θ	relative type I threshold
$\tilde{d}'_{\text{b}}$	meta-d'-balance
$\tilde{d}'_{\text{SSE}}$	meta-d'-SSE
tildes	meta-quantities

Table 2

Table of type I and II hit and false alarm rates for the example simulations at each level of metacognition

Level of metacognition	h	f	H_+	H_-	F_+	F_-
Low	0.7	0.35	0.6	0.6	0.5	0.5
Medium	0.7	0.35	0.6	0.6	0.4	0.4
High	0.7	0.35	0.7	0.7	0.35	0.35

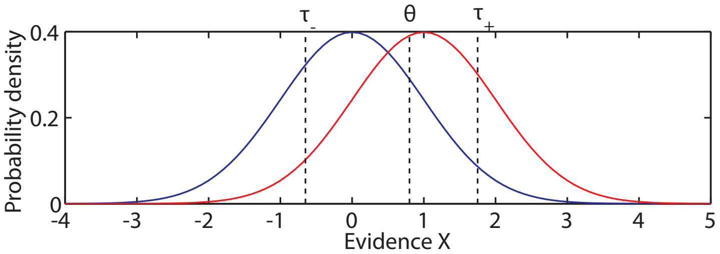

Figure 1. The ideal SDT model. The blue curve shows the distribution of the evidence X when the stimulus is absent ($S = 0$) and the red curve shows the distribution when the stimulus is present ($S = 1$). The stimulus is detected as present ($R = 1$) if X is greater than the type I threshold θ. Confidence is high ($C = 1$) if X is greater than the upper type II threshold τ_+ or less than the lower type II threshold τ_-. We define d' as the distance between means in units of the 'stimulus absent' distribution standard deviation (always 1 in our conventions). In this schematic $d' = 1$, the standard deviation of the 'stimulus present' distribution $\sigma = 1$, and the types I and II thresholds are set arbitrarily.

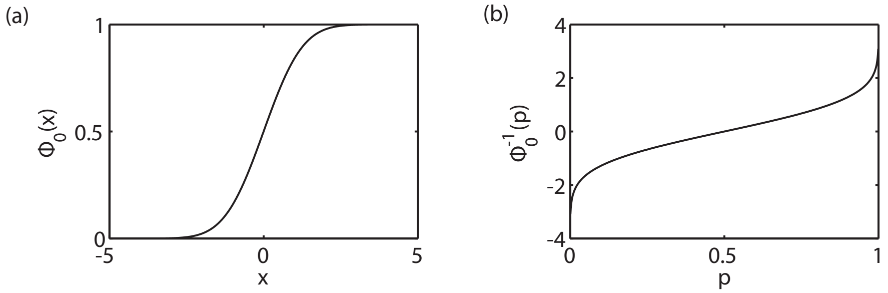

Figure 2. (a) The cumulative distribution function Φ_0 of the standard Gaussian distribution with mean 0 and standard deviation 1. (b) The inverse, Φ_0^{-1}, used in the formula (1) for d'. For small or large p, a small change in p leads to a large change in Φ_0^{-1}.

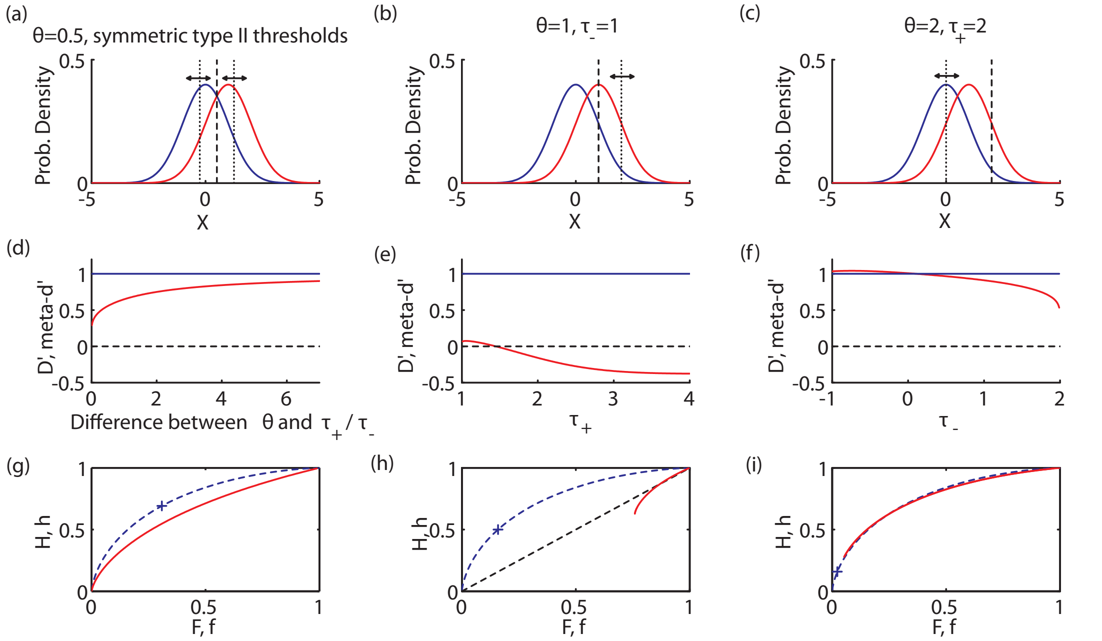

Figure 3. Type II d-prime (D') and meta-d' under varying decision and confidence thresholds for the standard SDT model with $d' = 1$ and $\sigma = 1$. Top row (a-c): Evidence distributions (red, present; blue, absent) and decision and confidence thresholds for (a) $\theta = 0.5$, τ_+ and τ_- symmetric about θ; (b) $\theta = 1$, $\tau_- = 1$, τ_+ variable; (c) $\theta = 2$, $\tau_+ = 2$, τ_- variable. Thick black arrows show the threshold being varied in each case. Middle row (d-f): D' (red) and meta-d' ($= \tilde{d}'_{\mathrm{b}} = \tilde{d}'_{\mathrm{SSE}}$) (blue) against varying type II thresholds, corresponding to the scenarios in (a-c) respectively. Dotted black lines mark where zero is on the vertical axis. Bottom row (g-i): Type II ROC curves (red) for H against F for the points plotted in (d-f) respectively. The blue cross in each panel shows the corresponding single type I ROC point, and the blue dashed line the full type I ROC curve. The dotted black line in (h) is the line $H = F$. By design, meta-d' is equal to d', for all choices of decision and confidence thresholds, and whichever variant of the measure is used. By contrast, D' is highly dependent on decision and confidence thresholds, and in extreme cases can be negative, or even greater than d'. Thus meta-d' but not D' is a stable measure of metacognition in these scenarios.

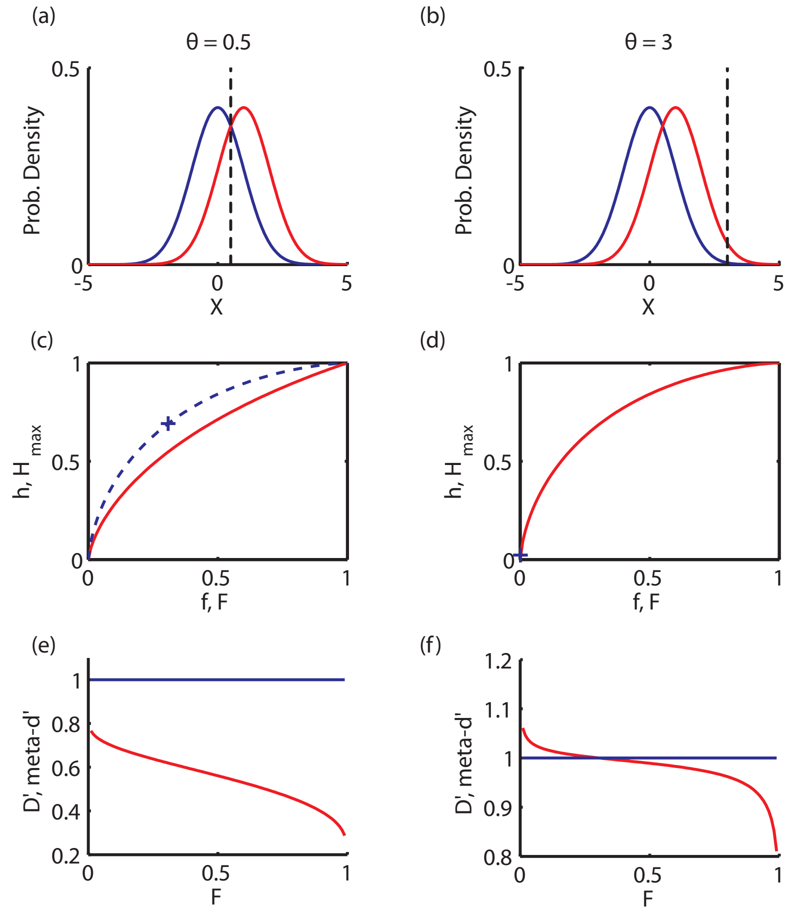

Figure 4. Optimal type II ROC curves, D' and meta-d' for two different choices of type I threshold. Top row (a,b): evidence distributions (red, present; blue, absent) and type I thresholds (dashed line, θ) for: (a) $d' = 1$, $\sigma = 1$, $\theta = 0.5$; (b) $d' = 1$, $\sigma = 1$, $\theta = 3$. Middle row (c,d): optimal ROC curves (red, H_{max} against F) corresponding respectively to (a) and (b). In these panels the blue cross indicates the corresponding single type I ROC point (f, h), and the blue dashed line shows the full type I ROC curve obtained by varying θ. Note that in (d) the type I and II ROC curves almost exactly coincide. Bottom row (e,f): D' (red) and meta-d' $(= \tilde{d}'_{\text{b}} = \tilde{d}'_{\text{SSE}})$ (blue) against F for the respective ROC curves in the row above (c,d). In each of the two cases, the optimal type II ROC curve compares differently with the corresponding type I ROC curve (c,d), and D' is highly variable (e,f). Only meta-d' remains constant, and therefore accurately reflects the optimal metacognition in both cases.

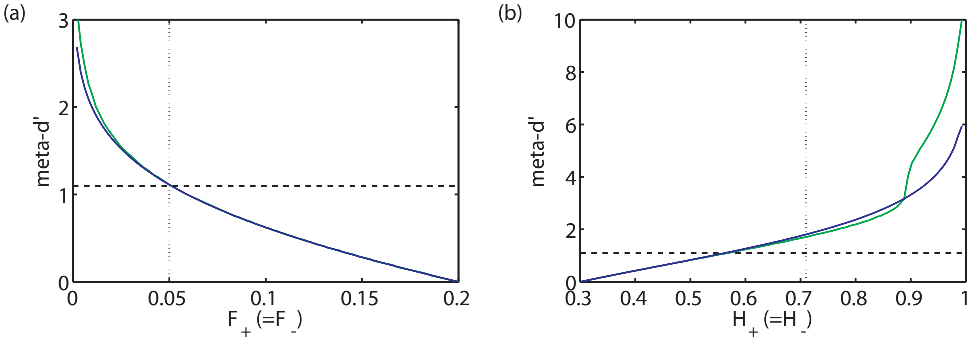

Figure 5. Behaviour of meta-d' under systematic variation of type II false alarm and hit rates. (a) $\tilde{d}'_{\mathrm{b}}$ (blue) and $\tilde{d}'_{\mathrm{SSE}}$ (green) against F_+ with $h = 0.8$, $f = 0.4$, $\sigma = 1$, H_+ and H_- fixed at 0.2, and F_+ and F_- fixed to be equal. (b) $\tilde{d}'_{\mathrm{b}}$ (blue) and $\tilde{d}'_{\mathrm{SSE}}$ (green) against H_+ with $h = 0.6$, $f = 0.2$, $\sigma = 1$, F_+ and F_- fixed at 0.3, and H_+ and H_- fixed to be equal. Dashed lines show d'. Dotted lines indicate the boundaries of the stable regions as defined by the criterion (29) for meta-d' measures to be stable. This excludes the region to the left in (a) and the region to the right in (b). For the stable regions, both $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$ give similar values for meta-d'.

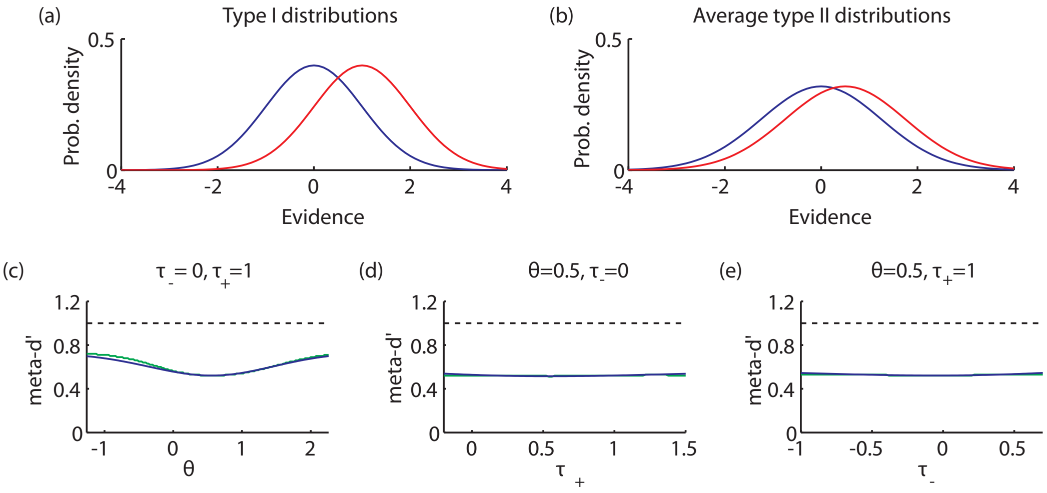

Figure 6. Behaviour of meta-d' on a degrading signal model. Top row: Evidence distributions for a degrading signal model (parameter values $d' = 1$, $\sigma = 1$, $a_0 = a_1 = s_0 = s_1 = 1/2$; see Supplemental Material for details). (a) Evidence distributions for the type I response. (b) Average evidence distributions for the type II response. Bottom row: Meta-d' for the distributions in (a,b) plotted against (c) type I threshold θ, (d) upper type II threshold τ_+ and (e) lower type II threshold τ_-, holding the other two thresholds constant in each case. Blue curves show $\tilde{d}'_{\mathrm{b}}$ and green curves show $\tilde{d}'_{\mathrm{SSE}}$. Dashed lines show the constant value of $d' = 1$. In each panel, the threshold is varied across the full range satisfying the inclusion criterion (29). Meta-d' values are approximately independent of decision and confidence thresholds (though less so for θ; see text) and are less than d', reflecting imperfect metacognition due to the degrading signal.

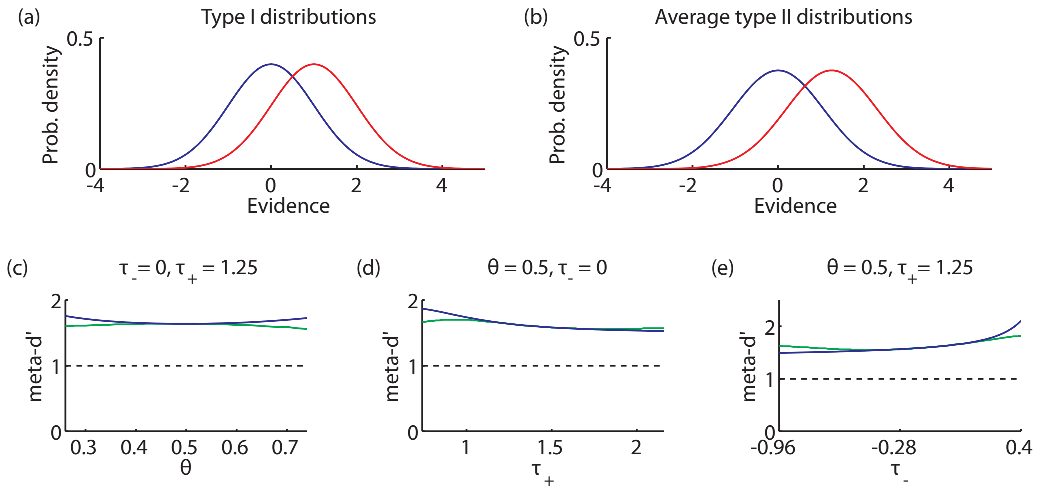

Figure 7. Behaviour of meta-d' on an enhancing signal model. Top row: Evidence distributions for an enhancing signal model with (parameter values $d' = 1$, $\sigma = 1$, $b_0 = b_1 = 1/4$; see Supplemental Material for details). (a) Evidence distributions for the type I response. (b) Average evidence distributions for the type II response. Bottom row: Meta-d' for the distributions in (a,b) plotted against (c) type I threshold θ, (d) upper type II threshold τ_+ and (e) lower type II threshold τ_-, holding the other two thresholds constant in each case. Blue curves show $\tilde{d}'_{\mathrm{b}}$ and green curves show $\tilde{d}'_{\mathrm{SSE}}$. Dashed lines show the constant value of $d' = 1$. In each panel, the threshold is varied across the full range satisfying the inclusion criterion (29). Meta-d' values are approximately independent of decision and confidence thresholds, and are greater than d', reflecting the enhanced evidence available for the type II metacognitive task.

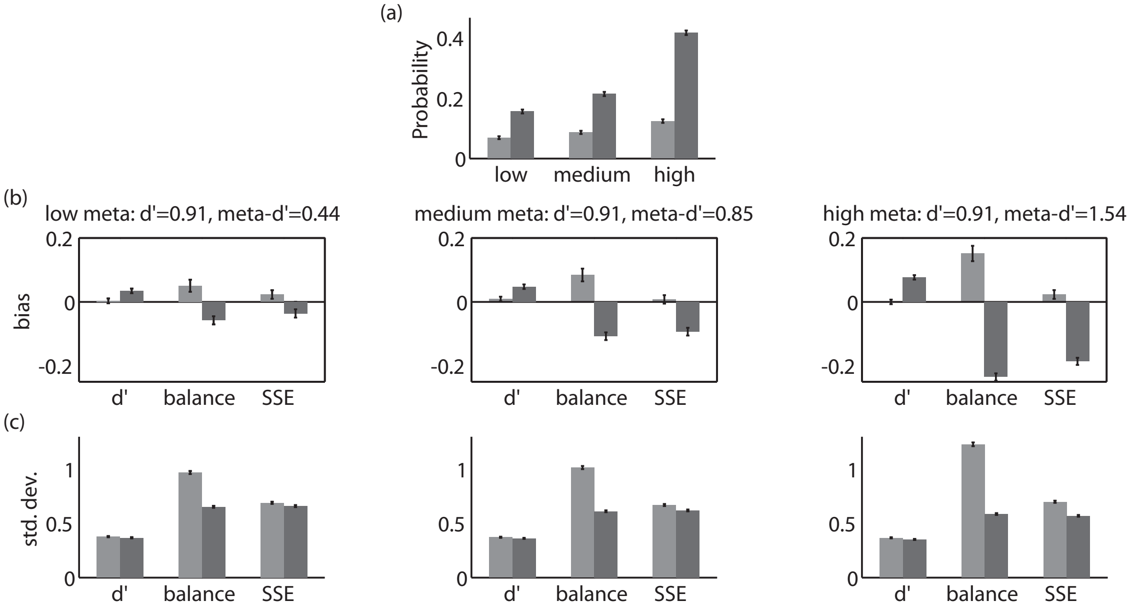

Figure 8. Bias and variance of d' and meta-d' in simulations of 10,000 (non-excluded) subjects, each performing 50 trials. (a) Probability of a subject being excluded in the low, medium and high metacognition examples for narrow (light grey) and wide (dark grey) exclusion criteria. (b) Bias and (c) standard deviation of d', $\tilde{d}'_{\mathrm{b}}$ (balance) and $\tilde{d}'_{\mathrm{SSE}}$ (SSE) for low, medium and high metacognition examples with narrow (light grey) and wide (dark grey) exclusion criteria. The hit and false alarm rates used in each of the simulations are given in Table 2. Error bars indicate 95% confidence intervals. Wide exclusion criteria lead to more bias but less variance in empirical meta-d'.

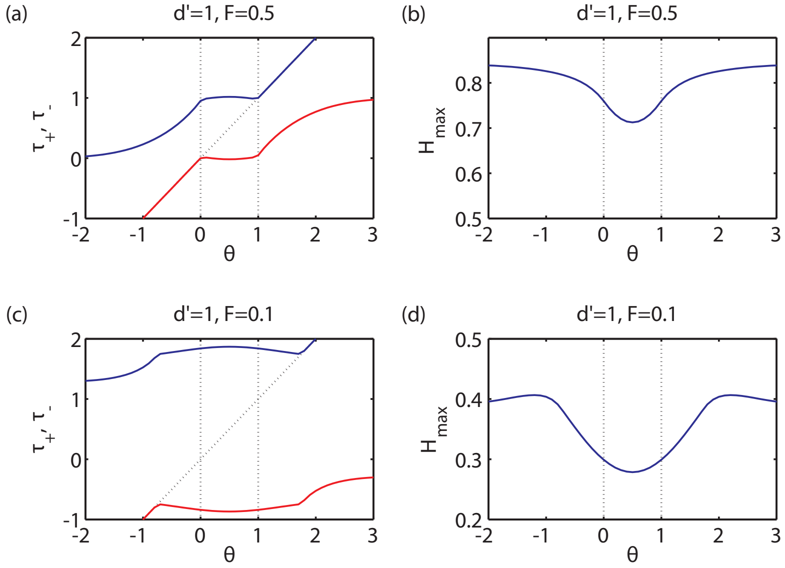

Figure 9. (a) Optimal τ_+ and τ_- and (b) $H_{\max}$ for $d' = 1$ and $F = 0.5$. (c) Optimal τ_+ and τ_- and (d) $H_{\max}$ for $d' = 1$ and $F = 0.1$. $\sigma = 1$ in both cases. Dotted vertical lines show where the peaks of the $S = 0$ and $S = 1$ distributions lie. The dotted diagonal lines in (a) and (c) show $\tau = \theta$. $H_{\max}$ shows substantial variability under changes in type I response bias and reaches a peak for type I threshold values that correspond to very strong type I response bias.

Further details and results on meta-d' on alternative SDT models

(Supplemental material for "Measures of metacognition on signal-detection theoretic models")

Adam B. Barrett, Zoltan Dienes and Anil K. Seth
University of Sussex

Mathematical description of degrading signal model

The general mathematical description of the type II evidence on this model is as follows. The degraded type II evidence when the stimulus is absent is

$$X_0^{(\mathrm{II})} \sim \mathcal{N}(a_0 x_0, s_0^2)\,, \tag{1}$$

where x_0 is the outcome of the type I evidence, and $0 < a_0 < 1$ and s_0 are free parameters. Similarly, when the stimulus is present the degraded type II evidence is

$$X_1^{(\mathrm{II})} \sim \mathcal{N}(a_1 x_1, s_1^2)\,. \tag{2}$$

We denote the type I threshold by θ and the type II thresholds by $\tau_\pm$ as above, but note that due to the degradation of the signal, the constraint $\tau_- < \theta < \tau_+$ is not needed. The type II hit rate for positive responses is then given by

$$\begin{aligned} H_+ &= P(X_1^{(\mathrm{II})} > \tau_+ | X_1 > \theta) && (3)\\ &= \frac{1}{h}\int_\theta^\infty P(X_1^{(\mathrm{II})} > \tau_+ | X_1 = x)\cdot P_{X_1}(x)\mathrm{d}x && (4)\\ &= 1 - \frac{1}{h}\int_\theta^\infty \phi_{d',\sigma}(x)\Phi_{a_1 x, s_1}(\tau_+)\mathrm{d}x\,. && (5)\end{aligned}$$

Similarly,

$$\begin{aligned} F_+ &= 1 - \frac{1}{f}\int_\theta^\infty \phi_0(x)\Phi_{a_0 x, s_0}(\tau_+)\mathrm{d}x\,, && (6)\\ H_- &= \frac{1}{1-f}\int_{-\infty}^\theta \phi_0(x)\Phi_{a_0 x, s_0}(\tau_-)\mathrm{d}x\,, && (7)\\ F_- &= \frac{1}{1-h}\int_{-\infty}^\theta \phi_{d',\sigma}(x)\Phi_{a_1 x, s_1}(\tau_-)\mathrm{d}x\,. && (8)\end{aligned}$$

Mathematical description of enhancing signal model

The general mathematical description of the type II evidence on this model is as follows. When the stimulus is absent, the type II evidence is given by

$$X_0^{(\mathrm{II})} \sim \mathcal{N}(x_0, b_0^2)\,, \tag{9}$$

where x_0 is the outcome of the type I evidence, and b_0 is a free parameter. Thus, some additional variance is added, reflecting an increase in noise, but the evidence remains the same on average. When the stimulus is present, the enhanced type II evidence is given by

$$X_1^{(\mathrm{II})} \sim \mathcal{N}(x_1 + b_1 d', b_1^2\sigma^2)\,, \tag{10}$$

where x_1 is the outcome of the type I evidence, and b_1 is a free parameter. The type II hit rates and false alarm rates are computed similarly to on the degrading signal model, such that

$$\begin{aligned} H_+ &= 1 - \frac{1}{h}\int_\theta^\infty \phi_{d',\sigma}(x)\Phi_{x+b_1 d', b_1\sigma}(\tau_+)\mathrm{d}x\,, && (11)\\ F_+ &= 1 - \frac{1}{f}\int_\theta^\infty \phi_0(x)\Phi_{x,b_0}(\tau_+)\mathrm{d}x\,, && (12)\\ H_- &= \frac{1}{1-f}\int_{-\infty}^\theta \phi_0(x)\Phi_{x,b_0}(\tau_-)\mathrm{d}x\,, && (13)\\ F_- &= \frac{1}{1-h}\int_{-\infty}^\theta \phi_{d',\sigma}(x)\Phi_{x+b_1 d', b_1\sigma}(\tau_-)\mathrm{d}x\,. && (14) \end{aligned}$$

Examples with unequal variances

Here we illustrate the behaviour of meta-d' measures on degrading and enhancing signal models with type I evidence distributions of unequal variance ($\sigma = 2$). Figure S1 shows behaviour on the degrading signal model and Figure S2 illustrates the enhancing signal model. These figures correspond to Figures 6 and 7 for the equal variance case.

Model with type I criterion jitter

SDT models often assume that the decision threshold remains stable over time; however it has been argued that trial-to-trial jitter in the decision threshold may exist (Ashby & Maddox, 1993; Mueller & Weidemann, 2008; Benjamin, Diaz, & Wee, 2009). Here we examine a model with type I criterion jitter to test whether this affects the independence of meta-d' from types I and II response bias. On this model, the type I and type II evidence are generated following the standard SDT model. However, while the type II thresholds $\tau_\pm$ remain constant across trials, the type I threshold is jittered according to an independent Gaussian random variable on each trial.

The mathematical description of this model is as follows. We denote the jittered type I threshold by $\Theta \sim \mathcal{N}(\theta, \eta^2)$. We denote the distance between 'present' and 'absent' distributions by d, dropping the prime since the actual d' [as measured by performance according to (1)] is affected by the jitter

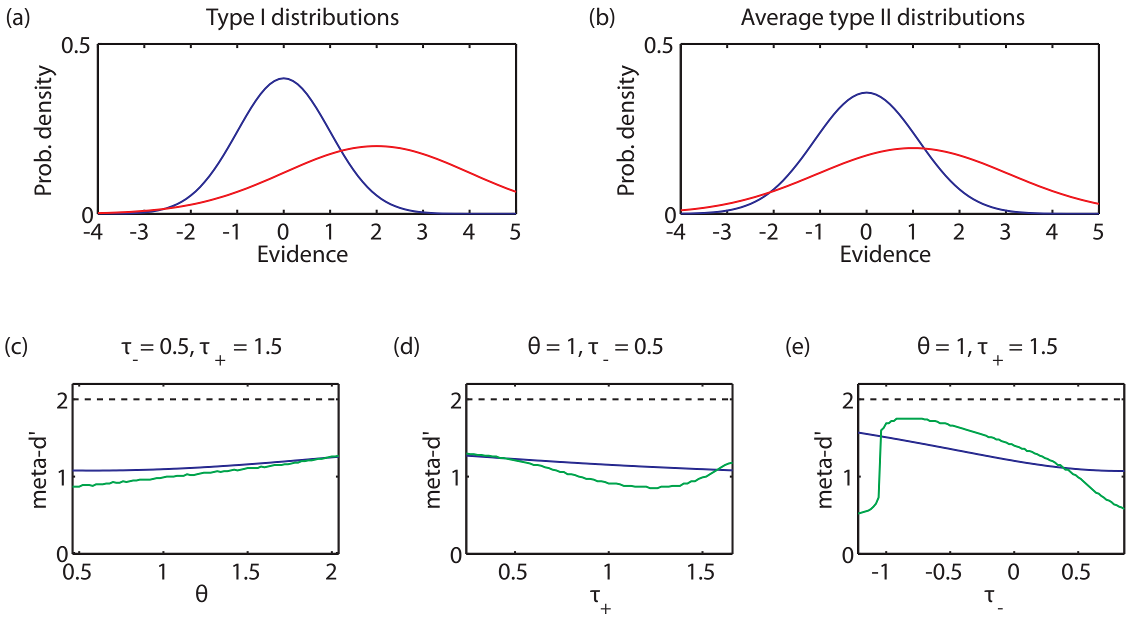

Figure S1. Meta-d' on a degrading signal model with unequal variances ($a_0 = a_1 = s_0 = s_1 = 1/2$, $d' = 2$, $\sigma = 2$). Top row: evidence distributions for (a) the type I response and (b) the type II response; stimulus absent in blue, stimulus present in red. Bottom row: behaviour of meta-d' for varying (c) θ, (d) τ_+, and (e) τ_-. Blue curves show $\tilde{d}'_{\mathrm{b}}$ and green curves show $\tilde{d}'_{\mathrm{SSE}}$. Dashed lines show the constant value of $d' = 2$. In each panel, the threshold being varied is taken across the full range that satisfies the inclusion criterion (29).

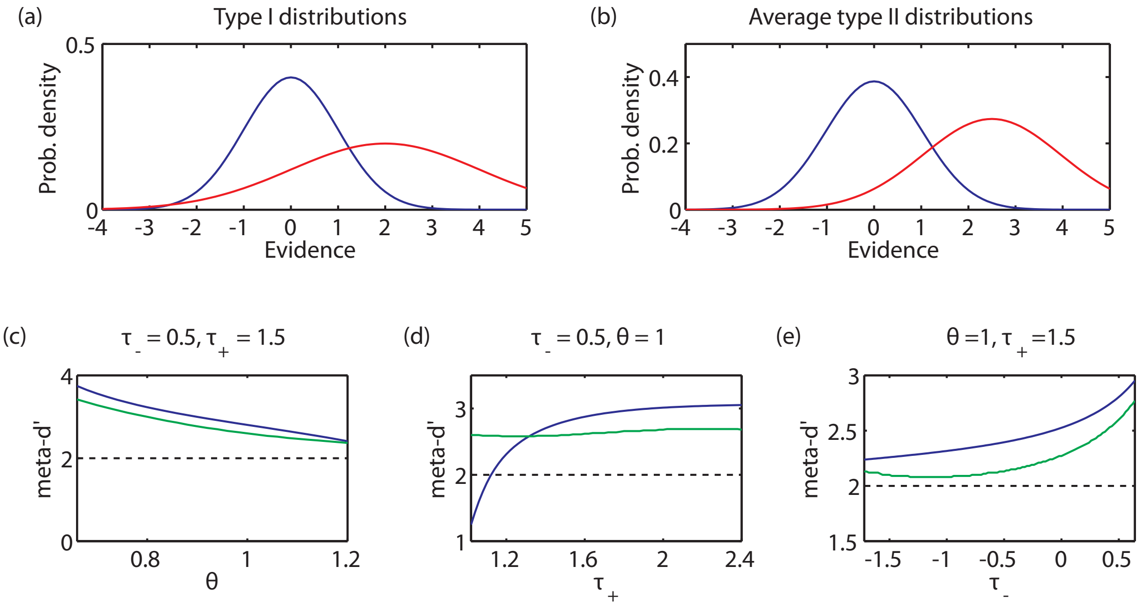

Figure S2. Meta-d' on an enhancing signal model with unequal variances ($b_0 = b_1 = 1/4$, $d' = 2$, $\sigma = 2$). Top row: evidence distributions for (a) the type I response and (b) the type II response; stimulus absent in blue, stimulus present in red. Bottom row: behaviour of meta-d' for varying (c) θ, (d) τ_+, and (e) τ_-. Blue curves show $\tilde{d}'_{\mathrm{b}}$ and green curves show $\tilde{d}'_{\mathrm{SSE}}$. Dashed lines show the constant value of $d' = 2$. In each panel, the threshold being varied is taken across the full range that satisfies the inclusion criterion (29).

and is less than d. It can be shown that the type I hit rate and false alarm rate are given by

$$\begin{aligned} h &= 1 - \Phi_0\left(-\frac{d-\theta}{\sqrt{\sigma^2+\eta^2}}\right), \qquad (15) \\ f &= 1 - \Phi_0\left(\frac{\theta}{\sqrt{1+\eta^2}}\right), \qquad (16) \end{aligned}$$

and hence

$$d' = \sigma\frac{d-\theta}{\sqrt{\sigma^2+\eta^2}} + \frac{\theta}{\sqrt{1+\eta^2}}\,. \qquad (17)$$

The type II quantities are derived as follows:

$$\begin{aligned} H_+ = &= P(X_1 > \tau_+ | X_1 > \Theta) \qquad (18) \\ &= \int_{-\infty}^{\infty} \mathrm{d}\theta' P(X_1 > \tau_+ | X_1 > \theta) P_\Theta(\theta') \qquad (19) \\ &= \int_{-\infty}^{\tau_+} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta') \frac{1-\Phi_{d,\sigma}(\tau_+)}{1-\Phi_{d,\sigma}(\theta')} + \int_{\tau_+}^{\infty} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta')\,, \qquad (20) \end{aligned}$$

and similarly

$$\begin{aligned} F_+ &= \int_{-\infty}^{\tau_+} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta') \frac{1-\Phi_0(\tau_+)}{1-\Phi_0(\theta')} + \int_{\tau_+}^{\infty} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta')\,, \qquad (21) \\ H_- &= \int_{\tau_-}^{\infty} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta') \frac{\Phi_0(\tau_-)}{\Phi_0(\theta')} + \int_{-\infty}^{\tau_-} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta')\,, \qquad (22) \\ F_- &= \int_{\tau_-}^{\infty} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta') \frac{\Phi_{d,\sigma}(\tau_-)}{\Phi_{d,\sigma}(\theta')} + \int_{-\infty}^{\tau_-} \mathrm{d}\theta' \Phi_{\theta,\eta}(\theta')\,. \qquad (23) \end{aligned}$$

Figures S3 and S4 show the behaviour of d', $\tilde{d}'_{\mathrm{b}}$ and $\tilde{d}'_{\mathrm{SSE}}$ on example criterion jitter models with respectively equal and unequal variances. In both examples d' is approximately independent of decision and confidence thresholds, and only slightly less than the distance d between the two evidence distributions. The meta-d' measures are approximately equal to d' for all decision and confidence threshold values, reflecting well the fact that there is no enhancement or degradation of the evidence in between the type I and II responses.

References

Ashby, F., & Maddox, W. (1993). Relations between prototype, exemplar, and decision bound models of categorization. Journal of Mathematical Psychology, 37(3), 372 - 400. doi: 10.1006/jmps.1993.1023

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. Psychological Review, 116(1), 84115. doi: 10.1037/a0014351

Mueller, S., & Weidemann, C. (2008). Decision noise: An explanation for observed violations of signal detection theory. Psychonomic bulletin and review, 15(3), 465494.

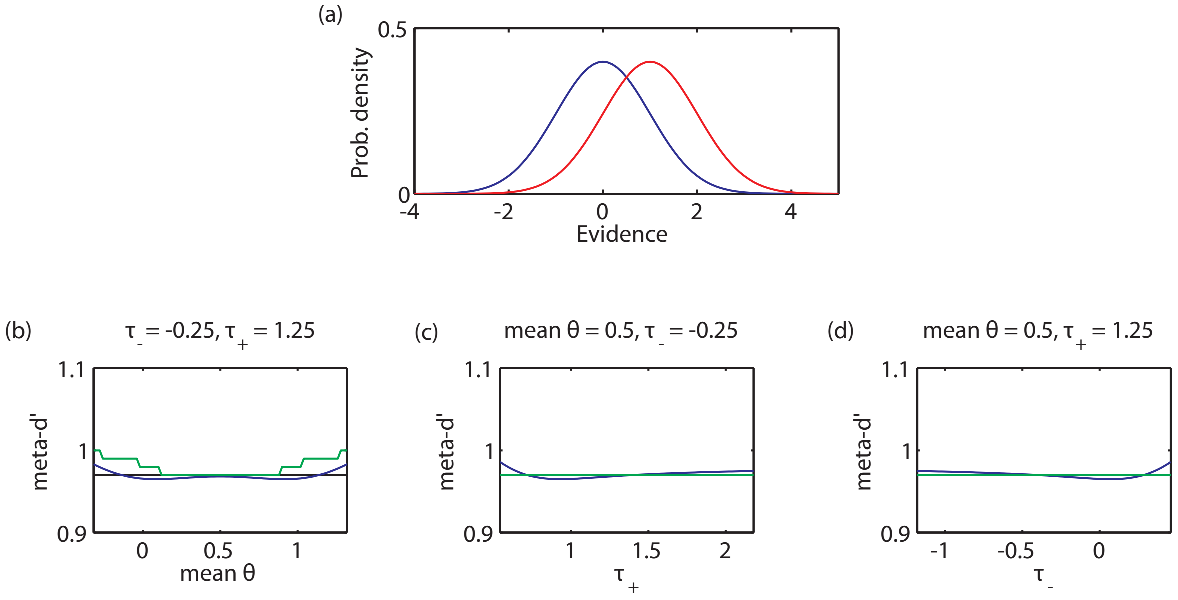

Figure S3. Meta-d' on a model with type I criterion jitter and equal variances ($d = 1$, $\sigma = 1$, $\eta = 0.25$). (a) Evidence distributions for the type I and II responses; stimulus absent in blue, and stimulus present in red. Bottom row: meta-d' for varying (b) mean decision threshold θ, (c) upper confidence threshold τ_+, and (d) lower confidence threshold τ_-. Blue curves show $\tilde{d}'_{\mathrm{b}}$, green curves show $\tilde{d}'_{\mathrm{SSE}}$. The black line in (b) shows d', which varies in this case due to the jitter. In (c) and (d) $d' = \tilde{d}'_{\mathrm{SSE}}$. In (b-d) the threshold is varied across the full range satisfying the inclusion criterion (29). While the jitter causes d' to be slightly reduced as compared to the distance between the two evidence distributions, meta-d' remains approximately equal to d'.

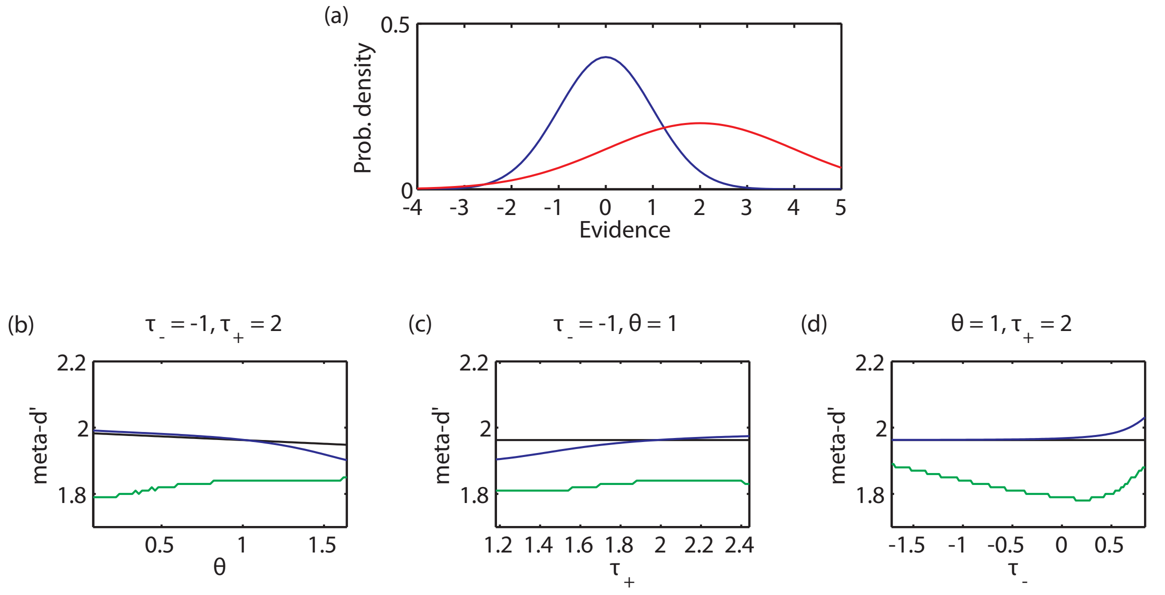

Figure S4. Meta-d' on a model with type I criterion jitter and unequal variances ($d = 2$, $\sigma = 2$, $\eta = 0.25$). (a) Evidence distributions for the type I and II responses, stimulus absent in blue and stimulus present in red. Bottom row: behaviour of meta-d' for varying (b) θ, (c) τ_+, and (d) τ_-. Blue curves show $\tilde{d}'_{\mathrm{b}}$, green curves show $\tilde{d}'_{\mathrm{SSE}}$, and black curves show d', which varies in this case due to the jitter. In each panel, the threshold being varied is taken across the full range that satisfies the inclusion criteria.