

# Inferences about Consciousness Using Subjective Reports of Confidence

Maxine Sherman, Adam B. Barrett and Ryota Kanai

## Chapter Abstract

Metacognition, or "knowing that you know", is a core component of consciousness. Insight into a perceptual or conceptual decision permits us to infer perceptual or conscious knowledge underlying that decision. However when assessing metacognitive performance care must be taken to avoid confounds from decisional and/or confidence biases. There has recently been substantial progress in this area and there now exist promising approaches. In this chapter we introduce type I and II signal detection theory (SDT), and describe and evaluate signal detection theoretic measures of metacognition. We discuss practicalities for empirical research with these measures, for example, alternative methods of transforming extreme data scores and of collecting confidence ratings, with the aim of encouraging the use of SDT in research on metacognition. We conclude by discussing metacognition in the context of consciousness.

## Keywords

Metacognition; consciousness; signal detection; type 2; confidence; SDT; visual awareness

## Introduction

An important aspect of consciousness is the ability to reflect upon one's own thoughts, an insight which can be traced back to John Locke, who stated that "*consciousness is the perception of what passes in a man's own mind*" (Locke, 1700). This definition of consciousness forms the basis of Higher Order Thought (HOT) theories of (phenomenal) consciousness (Gennaro 2004; Lau & Rosenthal 2011; Rosenthal 1986), which posit that it is those states for which we have some representation or conceptualisation that we have phenomenology of. It is not necessary to subscribe to this account of consciousness, however, to appreciate that our ability to reflect upon our own thoughts and decisions taps into an important facet of awareness. We can operationalise this ability as metacognitive sensitivity, performance or accuracy, terms used interchangeably<sup>1</sup>. These are defined as the ability to accurately judge the correctness of one's own decisions. We say that metacognitive sensitivity is high when confidence in the decision tracks task performance, and that it is low when confidence does not. It is the measurement of this 'tracking' that will form the subject of this chapter.

For reasons described above, metacognitive sensitivity is interesting in itself. It is also a valuable tool for measuring (indirectly) the extent to which a percept or knowledge is conscious. The argument, following from HOT, goes as follows: If task performance is above chance, there must be perception at least at the unconscious level. If indeed this performance is unconscious, then overall, participants should not be confident in their responses but should feel as though they were guessing. On the other hand, if the participant is conscious of the stimulus then they should be confident in their (correct) responses. In the case of stimulus detection this is more intuitive for trials where a target is present than when it is absent, yet we make this claim for both cases. The argument applies in the same way for measuring conscious knowledge, in which unconscious knowledge may generate Feelings of Knowing (Koriat 1995; Nelson 1984) or familiarity (Dienes, Scott, & Seth 2010; Dienes & Scott 2005). Thus we can use metacognitive performance as a proxy measure of awareness (Seth, Dienes, Cleeremans, Overgaard, & Pessoa 2008) although with caution (for debates in the literature, see Kunimoto, Miller, & Pashler 2001; Persaud, McLeod, & Cowey 2007; Rounis, Maniscalco, Rothwell, Passingham, & Lau 2010; Seth 2008; Song et al. 2011)

We will begin the chapter with a brief overview of type I Signal Detection Theory (SDT) which is often used to calculate objective task performance and which will be referred to regularly throughout the chapter. For a more thorough account of SDT we recommend Macmillan & Creelman (2004) and Green & Swets (Green & Swets 1966) . We will also cover ways in which the researcher may want to measure confidence and what we ultimately need for a good metacognitive measure. We will then move to a discussion of, largely, type II SDT measures of metacognition. These quantify metacognitive performance by examining the correspondence between the type I decision accuracy and confidence. Specifically, we will cover Pearson's  $r$ , the phi correlation coefficient and the Goodman-Kruskal Gamma coefficient in the first section, followed by type II  $D'$ , type II ROC curves, meta- $d'$  and meta- $d'$  balance. These will be discussed from the user's perspective and therefore cover their assumptions, their calculation and their respective advantages and caveats.

## **Measuring metacognition: Precursors**

### ***Type I Signal Detection Theory***

[Figure 1 here]

Signal detection theory (Green & Swets 1966; Macmillan & Creelman 2004) models the way in which we make binary choice perceptual decisions. Under the model, decision processes are inherently noisy. The choice is therefore between attributing the stimulation to just noise and attributing it to signal as well as noise. Alternatively, it considers the choice between a noisy 'type A' signal versus a noisy 'type B' signal. Here we will mainly consider the 'absent' versus 'present' scenario. However, all the methods work equally well for 'A' versus 'B' - 'A' can simply be substituted for 'absent' and 'B' for 'present'. The model is illustrated in figure 1. It is assumed that the probabilities of the stimulation being caused by noise and being caused by a noisy signal can each be modelled as Gaussian distributions along a continuous decision axis, often stimulation strength, for example, stimulus contrast. It is assumed that evidence accumulates along the decision axis and, depending on whether or not a certain threshold is reached, the stimulus is classified respectively as 'present' or 'absent'. This so-called decision threshold is modelled as a horizontal intercept called  $c$  or  $\theta$ . In yes/no tasks, this threshold is often expressed as  $\beta$ , which represents the likelihood of obtaining that

signal strength on a signal relative to noise trial. A  $\beta$  of 1 represents a bias-free observer.  $\beta$  greater than 1 represents a bias towards reporting noise, whereas  $\beta$  less than one represents a bias towards reporting a signal. Similarly, when considering decision threshold  $c$ , an unbiased observer has their criterion where the noise and signal plus noise distributions intersect (given equal values). A decision threshold greater than this is called "conservative", and one less than this, "liberal".

[Table 1 here]

Detection sensitivity  $d'$  is defined as the difference between the means of the noise and the signal plus noise distributions, in units of the standard deviation of the noise distribution. If the assumptions of SDT are met,  $d'$  will be invariant to decision bias. The first assumption SDT makes is that the variance of the signal and the noise distributions are equal. The second is that both the signal and the noise distributions are indeed Gaussian. It is the first of these assumptions which tends to be problematic in psychology research; however if an unequal variances model fits better, then the corrected  $d'_a$  can be used instead. For example, yes/no tasks are thought of as being fit best by an unequal variances model.

In order to calculate  $d'$  the researcher collects data in a 2 x 2 design such that a signal is present or absent and the participant can be correct or incorrect. This leads to a table of response variables as shown in table 1. We can then calculate the following:

$$\text{Hit rate} = \frac{\sum \text{Hits}}{\sum (\text{Hits} + \text{Misses})}$$

and

$$\text{False alarm rate} = \frac{\sum \text{False alarms}}{\sum (\text{False alarms} + \text{Correct rejections})}$$

Task performance  $d'$  can then be calculated as  $\phi^{-1}(h) - \phi^{-1}(f)$ , where  $h$  is the hit rate and  $f$  is the false alarm rate, and the decision threshold  $c$  can be calculated as

$$c = \frac{-\phi(h) - \phi(f)}{2},$$

where  $\phi^{-1}$  is the inverse cumulative probability density function of the standard Gaussian distribution of mean 0 and standard deviation 1 (also commonly known as the Z-statistic). These statistics are in the units of the standard deviation of the noise distribution when its mean is set to zero).

If the researcher is assuming that the variance of the noise and the signal-plus-noise distributions are unequal then a corrected  $d'$  can be calculated as

$$d'_a = s\phi^{-1}(h) - \phi^{-1}(f)$$

where  $s$  is the ratio of the standard deviation of the signal-plus-noise distribution to that of the noise distribution. To estimate  $s$  from data, confidence ratings must be collected in order to obtain hit and false alarm rates for multiple decision thresholds (as described below in the section 'Type I ROC curves'). Subsequently,  $s$  and  $d'_a$  can be computed from the best fit values for the above equation for all values of  $h$  and  $f$ . It is important to note that one can only assume that  $d'_a$  is (approximately)

invariant to changes in decision threshold  $c$  if one can assume that  $s$  has been estimated to a good degree of accuracy (Macmillan & Creelman 2004).

### *Transforming data with zero or one hit rate or false alarm rate*

There are occasions when one obtains hit rates or false alarm rates of zero or one. In these cases, data have to be transformed to avoid infinities in the equation for  $d'$ . These arise from the inverse-phi function going to plus/minus infinity at  $1/0$ . For  $d'$  to be finite, the hit and false alarm rates always lie strictly between 0 and 1.

In most cases, these situations can be avoided by ensuring that one collects a large number of trials per condition (at least 50) and that manipulations which may affect the decision criterion, for example performance-related reward or punishment, are not too strong. However in the case that extreme data are obtained, two main transformations exist in the literature. The first is one that adapts only those data which are problematic. Here, in an experimental set-up with  $n$  signal trials and  $(N-n)$  noise trials, a zero hit or false alarm rate would be replaced with  $1/2n$  or  $1/2(N-n)$  respectively. A hit or false alarm rate equal to one would be replaced with  $1-(1/2n)$  or  $1-(1/2(N-n))$  respectively. Thus each of these variables is transformed proportionately to the number of trials across which it is computed. For example, in the case that 25% of 100 trials are signal trials, a 0 or 1 hit rate would be shifted by  $1/50$  and a 0 or 1 false alarm rate by  $1/150$ . This method is called the  $1/2N$  rule (Macmillan & Kaplan 1985). An alternative transformation, the log-linear transformation, was proposed by Snodgrass & Corwin (1988). Here, all data cells (total hits, false alarms, correct rejections and misses), regardless of whether they are problematic or not, have 0.5 added to them. This is advantageous in that all data are treated equally and renders it impossible to have zero or one hit or false alarm rates. Moreover, this correction can be considered a (Bayesian) unit prior, in which the prior hypothesis is that  $d'$  and  $c$  are equal to zero (Barrett, Dienes, & Seth 2013; Meador & Dienes 2013). Hautus (1995) modelled the effects of both of these transformations on  $d'$  using Monte-Carlo simulated data sets. They found that both transformations can bias  $d'$  measures, and that while the log-linear rule systematically underestimated  $d'$ , the  $1/2N$  rule was more biased and could distort  $d'$  in either direction. Therefore although the log-linear rule is recommended over its counterpart, both in principle and practically, ideally data would be collected such that the risk of obtaining troublesome data are minimised. An evaluation of numerous alternative transformations by Brown & White (2005) concluded the same as Hautus.

### *Type I ROC curves*

Another approach to computing detection sensitivity is to create a receiver operating characteristic (ROC) curve, the area under which gives us detection sensitivity  $A_z$ . This method requires participants to give a rating response about stimulus class ( $S_1$ , versus  $S_2$ ), for example from 1 = definitely  $S_1$  to 6 = definitely  $S_2$ . A benefit of this method is that it can be implemented in 2-interval forced-choice (2IFC) tasks or other paradigms which do not explicitly generate hits, misses, false alarms and correct rejections, as the researcher plots hit rate against false alarm rate) via hypothetical decision criteria based on different thresholds of the responses. In order to plot an ROC curve where a response scale of length  $n$  has been used there will be  $n-1$  ways to partition responses into hypothetical levels of decision criterion. Each partition determines the boundary between  $S_1$  and  $S_2$ . For example, first one would partition the data such that a rating of 1 indicates an  $S_1$  response and

a rating of 2-6 indicates an  $S_2$  response. Then one would partition such that a rating of 1 or 2 indicates an  $S_1$  response and 3 to 6 indicates  $S_2$ , continuing until a rating of 1-5 indicates an  $S_1$  response and a rating of 6 indicates an  $S_2$  response. Therefore, for each level of  $c$  (the partitioning) one obtains different numbers of hits and false alarms, and thus can compute the hit rate and false alarm rate for each. These are plotted against each other, producing a curve that characterizes sensitivity across a range of decision biases without making assumptions about the underlying signal and noise distributions. The diagonal on the graph represents chance performance, and the higher the curve lies above the diagonal, the greater the sensitivity, in that for any given false alarm rate the corresponding hit rate is higher. Thus the area under the ROC curve represents discrimination performance. This is easy to estimate from a basic plot of all the points obtainable from a dataset.

It should be noted that because it does not rely on the assumptions of SDT, ROC curve analysis is not technically SDT. Alternatively, if one does assume that decisions are made based on an SDT model (with not necessarily equal variances for the signal and signal plus noise distributions), then the Z-transform of the ROC curve is a straight line, and the area under the (non transformed) ROC curve can be obtained from a simple formula in terms of the slope and intercept of the Z-transform:

$$A_z = \Phi \left[ \frac{\text{Intercept}}{\sqrt{1 + \text{slope}^2}} \right]$$

A benefit of plotting an ROC curve or using SDT's  $d'$  is that as well as assessing whether objective task performance has changed following a manipulation, task performance can be decomposed into possible drivers of the change: hit rate and false alarm rate. For example, some empirical questions might hypothesise a change in  $h$  but not  $f$ . Kanai, Muggleton & Walsh (2008) found that transcranial magnetic stimulation over intraparietal sulcus induces perceptual fading by demonstrating such an asymmetry: Although overall detection performance reduced with TMS, only  $h$  and not  $f$  was affected. This pattern is consistent with the fading of a present target – an additional decrease in  $f$  would suggest that more general perceptual sensitivity had improved.

### ***Measuring metacognitive accuracy***

In order to assess the participant's judgement of their own accuracy the researcher needs to collect both an accuracy measure and a confidence measure. Typically, experimental designs include some objective task such as target detection or word recall in which objective performance can be measured. To measure metacognitive sensitivity we use what is known as 'the type II task', first coined by Clarke, Birdsall & Tanner (1959) and Pollack (1959), and so-called in reference to the aforementioned type I task of making decisions or judgements about the 'state of the world'. The type II task is to evaluate the accuracy of one's own decision. Galvin, Podd and Whitmore (2003) discuss the type II task and argue that

*"...The fact that the second decision [confidence that the trial was a signal trial] is a rating and follows a binary type I decision does not make it a type II decision. If the second decision is a rating of*

*confidence in the signal event rather than in the correctness of the first decision then it is a type I rating, no matter when it occurs."*

Following this, it is advised that the confidence judgement requested refers to the accuracy in the participant's decision. However from the perspective of consciousness science it seems counterintuitive to assume a distinction between asking for confidence in the signal and asking for confidence in the participant's judgement; this suggests an asymmetry in the trustworthiness of the objective (type I) and subjective (type II) responses. If we instead take type I decisions as those which are about the state of the world, then we can take type II decisions as probing the mental state or representation the subject has of the stimulus. In this sense the prompt "Confidence?" should be equivalent to the prompt "Confidence that you are correct?", though to our knowledge this has not been empirically addressed.

### *Collecting confidence ratings*

The traditional method of collecting confidence ratings is in two-steps: the judgement is made and then confidence is given, either in a binary fashion or on a scale. Whether confidence is collected on a scale or in a binary fashion will dictate the metacognitive measures available to use. Confidence scales (e.g. from 1 to 4) have the advantage of being more sensitive and they can later be collapsed into a binary scale, reducing the chance of getting 0 or 100% confident responses. However importantly, if conclusions about consciousness are to be drawn, we can only infer unconscious knowledge of perception from those trials where participants have reported no confidence (i.e. we cannot infer this from low confidence). Therefore a rating scale should only be symmetrically collapsed into a binary scale if no conclusions are to be drawn about awareness. If conclusions relate to metacognition then this would be fine. If the question of interest relates only to perceptual awareness, the Perceptual Awareness Scale could be used instead. This scale asks participants to rate the subjective visibility of their percept on a scale of one to four and is advisable for simple (e.g. stimulus detection) rather than complex (e.g. stimulus discrimination) designs (Dienes & Seth 2010; Sandberg et al. 2010), as the conscious content itself is not probed (Dienes & Seth, 2010). This is discussed in more detail in **[reference to Perceptual Awareness Scale book chapter]**.

Instead of requesting a type I and then a confidence response, report and confidence can be reported in a one-step procedure in which participants are asked to choose between two responses  $S_1$  and  $S_2$  and high and low confidence at the same time. For example, a rating scale could be used where the lowest value indicates high confidence in  $S_1$  and the highest value indicates high confidence in  $S_2$ . This has the benefit of being a faster reporting procedure. In the case of perceptual experiments it has been shown that, although one-step and two-step procedures generate different reaction times, they do not affect the confidence-accuracy correlation (Wilimzig & Fahle 2008). This, however, has not been verified for other type II measures. When using metacognition to assess the presence of conscious structural knowledge, one versus two step procedures do not tend to have an effect either. We refer the reader to Dienes (2008) and Wierzshon et al. (2012) for more detail.

### *What makes a good measure of metacognition?*

In order to assess the ability of an individual to monitor the accuracy of their decisions we need to be able to separate the information on which their decision is based from the insight into that information that they hold. Maniscalco and Lau (2012) articulate this well by distinguishing between

absolute and relative metacognitive sensitivity. Absolute metacognitive sensitivity refers to the relationship between confidence and accuracy alone, whereas relative sensitivity refers to the efficacy of the metacognitive evaluation without the confound of information quality. Therefore, in experiments where metacognitive performance is contrasted across two conditions it is imperative that objective performance is equated if one wants to measure relative rather than absolute metacognitive sensitivity.

In order to tap into relative metacognitive sensitivity we need a measure of how confidence tracks accuracy that is invariant to decision threshold (yes versus no) and confidence threshold (confident versus guess) or at least allows us to separate them. For example, by demonstrating reduced perceptual metacognition after theta-burst transcranial magnetic stimulation to prefrontal cortex, Rounis et al. (2010) were able to implicate this area in metacognitive sensitivity. They used bias-invariant (type II) meta- $d'$  (to be discussed later) as their measure, which allowed them to rule out the alternative interpretation that PFC is involved in determining confidence bias.

It is important to note that dependence on decisional or confidence biases is not problematic if one is aiming more simply to rate the subject's performance on the type II task. Viewed this way, metacognition may be facilitated *because* of shifts in confidence bias. Sherman et al. (under review) found that when perception of target presence or absence is congruent with a prior expectation, metacognition for the perceptual judgement improves. This result was successfully computationally modelled using an SDT model with decision and confidence criteria modulated by prior expectations.

Signal detection theoretic methods are useful because they allow us to consider the above points. By enabling the calculation of response and confidence biases as well as type I and II performance one can see how measures of task performance and decision bias interact. Further, one can see whether improvements in metacognitive performance can be attributed (at least in part) to specific changes in behaviour (for example, increased confidence for correct reports, called type II hit rate).

One also has to consider whether to obtain a single measure of metacognition across all trials, or whether to assess metacognition separately for each possible class of type I response, i.e. to use a so-called 'response-conditional' measure of metacognition. For example, in a target detection experiment, one has the classes "respond present" and "respond absent" (see table 3). Kanai, Walsh & Tseng (2010) defined the Subjective Discriminability Index (SDI) as a measure of subjective unawareness of stimuli, based on response-conditional type II ROC curves (see below). This can also be applied to type II  $D'$ . Specifically, by using only trials where subjects reported absence of a target (type I correct rejections and misses) in the type II calculation, we get a measure of metacognition for perception of absence. Their logic was that chance metacognitive accuracy implies blindness to the stimulus, whereas above chance metacognitive accuracy implies that although the subject reported the target as unseen, some perceptual awareness must have been present (inattentional blindness). This follows from participants' ability to appropriately modulate their post-decisional confidence according to their accuracy. The analogous 'Respond-present'-conditional measure does not seem to have an analogous interpretation to that of SDI, that is, in terms of visual consciousness. If participants are demonstrating above-chance detection performance but their metacognitive performance when they report target *presence* is at chance, it is perfectly possible that this may address our experimental hypothesis but it does not seem to be interpretable in and of itself.

With these points in mind we hope that we have given the reader a grasp of what is necessary from a paradigm in which metacognition is going to be measured. We will now continue and cover various statistics available in the literature.

## Confidence-Accuracy correlations

The most intuitive measure of metacognition would tell us whether accuracy and confidence are significantly and highly correlated. Two main alternatives are available: Pearson's  $r$  and phi. These are equal in the binary case, but distinct for the non-binary case (that is, if confidence is reported on a scale, the latter can be used).

For paired variables  $X$  and  $Y$  corresponding to confidence and accuracy values for  $n$  participants, the correlation  $r$  between confidence and accuracy is calculated as

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{s_x} \right) \left( \frac{Y_i - \bar{Y}}{s_y} \right)$$

where  $s_x$  and  $s_y$  are the sample standard deviations of  $X$  and  $Y$  respectively.

Alternatively, the phi correlation coefficient is calculated as

$$\phi^2 = \frac{\chi^2}{n}$$

where  $\chi^2$  is the chi-squared statistic and  $n$  is the number of participants.

When  $X$  and  $Y$  are binary, e.g.  $X$  equals 0 for low confidence and 1 for high confidence, and  $Y$  equals 0 for incorrect and 1 for correct, phi and  $r$  are equal to each other, and can be calculated from the

$$\phi = \frac{n_{1,1}n_{0,0} - n_{1,0}n_{0,1}}{\sqrt{n_{.,1}n_{.,0}n_{1,.}n_{0,.}}}$$

where  $n_{x,y}$  is the total number of trials on which  $X = x$  and  $Y = y$ , and  $n_{.,y}$  and  $n_{x,.}$  are respectively the total number of trials for which  $Y=y$  and  $X=x$ . Though simple, the problem with such a measure (and indeed, with any non-signal detection theoretic measure) is that  $r$  and  $\phi$  can be inflated by bias without there being a true improvement in metacognitive accuracy. To illustrate, we can imagine a stimulus detection paradigm whereby stimulus contrast is titrated such that performance is at 70% for all participants. However if one participant has a bias towards being confident whereas another tends to say they are guessing, the first of these participants will have a higher correlation between confidence and accuracy than the second without necessarily having increased insight into their own decision accuracy. Unfortunately these biases from criteria plague a lot of metacognitive measures.

## Goodman-Kruskal Gamma Coefficient

The Goodman-Kruskal Gamma coefficient,  $G$  (Goodman & Kruskal 1954) is a non-parametric analogue of the signal detection theoretic measure  $d'$ . Its appeal lies in its straightforward probabilistic operationalization, which overcomes problems surrounding assumptions about equal variance or normality. In its original form it is computed via the same  $2 \times 2$  factors as  $d'$  and it can be extended to cases in which ratings are given on a response plus confidence scale (e.g. 1 = very confident no, 6 = very confident yes). By being distribution-free it hoped to also be a flexible

measure of metacognitive accuracy when applied to type II data (Nelson 1984). Task performance  $V$  is characterised as follows for a 2 x 2 design, the construction aimed at eliminating dependence on overall response bias. Suppose there are two trials and one of them is 'stimulus present' and one of them is 'stimulus absent', and the subject responds 'present' on one trial and 'absent' on the other. Then  $V$  is the probability that these responses match the stimulus. The estimate of this (obtained from the data from all trials) is given by:

$$V = \frac{\Sigma \text{ hits} \times \Sigma \text{ correct rejections}}{(\Sigma \text{ hits} \times \Sigma \text{ correct rejections}) + (\Sigma \text{ misses} \times \Sigma \text{ false alarms})}$$

The Gamma coefficient is then given by

$$G = 2V - 1 = \frac{(\Sigma \text{ hits} \times \Sigma \text{ correct rejections}) - (\Sigma \text{ misses} \times \Sigma \text{ false alarms})}{(\Sigma \text{ hits} \times \Sigma \text{ correct rejections}) + (\Sigma \text{ misses} \times \Sigma \text{ false alarms})}$$

To assess metacognitive performance, pairs of responses (on the confidence scale) are combined to produce an analogue of  $V$ . There is no simple formula for the general (non 2 x 2) case, so for a thorough explanation we refer the reader to Masson & Rotello (2009).

In order to verify  $G$ 's supposed invariance to bias and distributional assumptions, Masson & Rotello (2009) simulated datasets in which metacognitive sensitivity was fixed and calculated  $G$ . More specifically, a 2AFC task was modelled as two probability distributions representing each choice. The difference between the means of these distributions was adjusted on simulation runs such that "population gamma", calculated by randomly sampling from the distributions in order to approximate the proportion of cases where  $A > B$ , was fixed. It was then compared to the Gamma obtained when considering decision biases. Indeed, they found that  $G$  actually does get distorted by decisional biases. Moreover, this distortion increased when data were simulated from an unequal variance model, suggesting that the invariance under reasonable changes to distributional assumptions may not hold.

## Type II $D'$

[Insert tables 2 and 3 here]

Type II signal detection theory extends the logic of its type I counterpart by using confidence reports to map onto detection accuracy (Kunimoto et al. 2001; Macmillan & Creelman 2004). It assumes that correct and incorrect responses can be plotted on a (type II) decision axis as Gaussian random variables, analogously to the signal and noise distributions in type I SDT. The distance between the peaks of the distributions gives us our measure of metacognitive sensitivity, type II  $D'$ .

As shown in tables 2 and 3, type II variables are computed analogously to type I variables, but instead of examining the correspondence between signal and response, response accuracy and confidence are compared. We define the type II hit as a confident and correct response, a type II false alarm as a confident but incorrect response, a type II miss as a correct but unconfident response and a type II correct rejection as an appropriately unconfident, incorrect response. Metacognitive performance type II  $D'$  is then calculated analogously to type I  $d'$  – by subtracting the normalised type II hit rate from the normalised type II false alarm rate. The type II decision criterion then represents confidence bias; the extent to which the subject is over or under-confident.

This measure generated much excitement when it was proposed by Kunimoto et al. (2001) as free of dependence on bias. Unfortunately, the biases were artificially fixed by the nature of the authors' paradigm; confidence was assessed by the magnitude of their wager on each trial, but the total wager they could place was fixed for each session. Their claim of invariance to confidence bias has since been found to neither hold empirically (Evans & Azzopardi 2007) nor theoretically (Barrett et al 2013) when type I and II decisions are made based on the same evidence. Indeed, Barrett et al. (2013) found that under certain circumstances (response-unconditional)  $D'$  is highly unstable. For example, if the type I criterion is placed where the noise and signal and noise distributions intersect then  $D'$  is maximised when the observer is maximally unconfident, which would be a nonsensical and mal-adaptive strategy. Barrett and colleagues also found by varying decision and confidence criteria that  $D'$  can range from being negative (which is difficult to interpret in a meaningful way) to being greater than type I  $d'$ . Importantly, these analyses demonstrate a high reliance of  $D'$  on decision and confidence thresholds. The behaviour of type II  $D'$ , then, does not suggest it to be a reliable measure of metacognition.

There are in fact problems with the general validity of the underlying statistical assumptions of type II  $D'$ . The assumption that type I sensory evidence is normally distributed is generally accepted because, by the central limit theorem, summed neural responses from a large population of neurons to targets will be Gaussian. However for the type II case it is less likely that the evidence for correct and incorrect responses can actually be represented as Gaussian distributions along a single decision axis (Maniscalco & Lau 2012). Indeed, Galvin et al. (2003) have demonstrated that the type I decision axis (e.g. signal strength) cannot be transformed in such a way that the type II evidence distributions for correct and incorrect decisions are Gaussian. Thus if a single pathway underlies both type I and II decisions then  $D'$  is not a measure of metacognition that arises naturally from SDT modelling.

Despite these problems, type II  $D'$  can still be useful as a basic measure of type II performance in some scenarios. Specifically, this is the case if response bias is small. Moreover, it is useful for situations when the number of trials per subject is small and more sophisticated measures, such as those based on area under an ROC curve, or meta- $d'$  (see below), are too noisy to yield significant results (Sherman et al, under review).

### **Type II SDT**

[Figure 2 here]

Type II  $D'$  is not the only way to envisage a type II signal detection theoretic model of metacognition. The transformation of the type I model into the type II domain can be achieved by considering the probability distributions for the evidence for signal absent and signal present, as per type I SDT. However we can overlay confidence thresholds  $\tau_+$  and  $\tau_-$  onto the type I decision axis, as illustrated in figure 3, such that confidence is high when  $X < \tau_-$  or  $X > \tau_+$  and low otherwise. This is illustrated in figure 3. Although this renders type  $D'$  an unprincipled measure, it invites certain promising alternatives, namely type II ROC curves and meta- $d'$ , as described below.

### **Type II ROC Curves**

While the type I ROC curve plots the probability of type I hits against the probability of type I false alarms for each level of criterion  $\theta$ , the type II ROC curve plots the probability, for some fixed type I

decision criterion  $c$ , type II hit rate against type II false alarm rate for all possible confidence thresholds. As they incorporate a range of thresholds they have been proposed to characterize metacognition in a stable manner. However, because at the type II level there are two thresholds,  $\tau_+$  and  $\tau_-$  (confidence thresholds for positive and negative type I responses respectively), for the response-unconditional case, three parameters are left to vary freely ( $\theta$ ,  $\tau_+$  and  $\tau_-$ ) rendering an infinite number of type II ROCs. Even if one fixes  $\theta$  at the empirically observed level the type II ROC is still not unique.

The current literature posits three potential solutions to this.

Galvin et al. (2003) suggested collapsing the two confidence thresholds into one likelihood function: the likelihood ratio of being correct versus incorrect. This enables a unique solution for fixed  $\theta$  and is straightforward to compute. However, the authors still found a strong dependence of the area under the curve (AUC) on  $\theta$ .

An alternative measure, proposed by Clifford et al. (2008) suggested that we could compare the type I ROC curve based on a confidence rating scale with the ROC curve obtained by manipulating  $c$  experimentally. That is, if we manipulate the physical properties of the stimulus such that response threshold changes (e.g. threshold contrast) then we can plot the false alarm rate/hit rate trade-off across artificially induced criterion shifts. This is the traditional type I ROC curve. We can compare this with an alternative type I ROC in which changes in criterion are modelled by differentially bisecting into "signal" and "noise", an  $n$  point rating scale all  $n-1$  ways. If metacognition is SDT-optimal, these two ROCs should be the same. This point follows from the assumption that an optimal observer would fully use the same information for the type I and the type II decision. Thus, Clifford et al. proposed their divergence as a measure of metacognition. Again though, the degree of divergence is not in general independent of type I response bias.

Finally, Barrett et al. (2013) constructed the SDT-optimal type II ROC curve; the type II ROC curve that, for fixed  $\theta$  and fixed type II false alarm rate ( $F$ ), gives us the greatest type II hit rate ( $H$ ),  $H_{\max}$  (and therefore type II performance). Similarly to the idea above, this describes the performance of the SDT-optimal observer. The algorithm for calculating  $H_{\max}$  is in the paper of Barrett et al. (2013). Unfortunately this curve was also found to be vulnerable to distortions from  $\theta$  however because it describes SDT-expected performance it can be used to check whether data conform to SDT.

The response-conditional case is more straightforward, as in that scenario, if one can ensure that the type I (decision) threshold is fixed, a unique type II ROC curve is obtained, by varying a single confidence threshold. The response-negative area under the ROC curve forms the basis for Kanai et al.'s (2010) SDI measure.

### **Meta-d' and Meta-d'-balance**

Meta-d' (Maniscalco & Lau 2012) and meta-d' balance (written here as meta-d'<sub>b</sub>; Barrett et al. 2013) are currently the gold standard in measures of metacognition. While type II D', as mentioned above, computes metacognitive sensitivity as a function of accuracy and confidence, meta-d' computes the (type I) accuracy that would be expected given the type II level information, if the observer were SDT-optimal. In this way, one can compare meta-d' to d' and assess metacognitive sensitivity relative to the SDT-ideal observer. The difference between meta-d' and d' has a clear interpretation in units

that correspond to the standard deviation of the noise distribution. Type II  $D'$  on the other hand is formulated in different units from type I  $d'$ , making it hard to directly compare these two measures.

If meta- $d'$  is equal to  $d'$  one assumes the participant has optimal metacognitive performance, whereas if it is lower than  $d'$ , one assumes that the optimal observer could achieve the empirical metacognitive performance with less type I information than the participant, rendering their performance suboptimal. It is assumed that meta- $d'$  will never be higher than  $d'$  as this would suggest the participant performed "super-optimally". In practice, this would support a model in which the observer has more information when making the type II decision than when making the type I decision, for example, after having had feedback on the type I decision or having had to make a speedy type I decision.

There are several possible operational definitions of meta- $d'$ , all of which rely on solving two pairs of equations, one pair obtained by considering type II performance following a positive type I response and the other obtained by considering type II performance following a negative type I response. All existing approaches fix the type I response bias (the relative type I threshold) to the observed value for the purposes of solving the equations for meta- $d'$ . The two pairs of equations cannot in general be solved simultaneously. Maniscalco & Lau (2012) adopt a data-driven approach, by proposing two methods for finding the best fit: minimizing the sum of the squares of the errors leads to meta- $d'_{SSE}$ , while maximum likelihood estimation leads to meta- $d'_{MLE}$ .

Barrett et al. (2013) introduced meta- $d$  balance (meta- $d'_b$ ) which, rather than assuming symmetry between positive and negative responses, permit response-conditional meta- $d'$  for positive and negative responses to differ. They propose this as a theory-driven rather than data-driven approach which affords an alternative calculation of meta- $d'$ . As with meta- $d'$ , they derive formulae for both positive and negative response-conditional meta- $d'$ , but rather than solving these simultaneously, they take their mean solution, weighted according to the number of positive relative to negative type I responses. Barrett et al. (2013) noted that the response-conditional meta- $d'$  measures do not on their own provide stable, response bias-invariant measures of metacognition; stability only comes when they are combined into a single measure.

Barrett et al. (2013) assessed how both meta- $d'_b$  and Maniscalco & Lau's meta- $d'_{SSE}$  behave under non-traditional SDT models. In practice, empirical data are messy and the paradigm may induce certain changes in how we envisage the statistical distributions of signal and noise. For example, Maniscalco & Lau (2012) write that if meta- $d'_{SSE}$  is being used, it would be preferable to utilise a 2IFC task than a target detection task because target detection tasks are generally modelled as an unequal variance model. Importantly, Barrett and colleagues found that under an unequal variance model, even when departing from standard SDT (i.e. when the signal is enhanced or degraded between the type I and II levels, or when type I criterion is jittered across trials, representing fluctuations in attention) both versions remain relatively robust, especially when the type I threshold is varied. In these cases, however, meta- $d'_b$  seems slightly more consistent than meta- $d'_{SSE}$ , which is unsurprising given that meta- $d'_b$  permits differences between the response-conditional metacognitive measures. Under signal-degradation, signal-enhancement and criterion jitter models, when variances are equal both measures were largely invariant to changes in type I and II thresholds.

Barrett et al. (2013) also looked at the behaviour of both meta- $d'$  measures on finite data sets and found that with small numbers of trials (approximately 50 trials per subject) both showed statistical bias and had higher variance than  $d'$ . However when 300 trials per subject were included in the analysis, bias approached zero and variance dropped substantially. Therefore to get the most out of these measures high numbers of trials per condition should be obtained.

The calculation of meta- $d'$  is optimal when no type I or II hit or false alarm rate is too extreme, and not possible when any of these take the value zero or one. This leaves one with two possible sets of data exclusion criteria to consider. The 'narrow exclusion criteria' only exclude a subject if any of the type I or response-conditional type II hit rates or false alarm rates are zero or one. These obviously maximise the number of subjects retained. An alternative choice is to use 'wide exclusion criteria' which exclude subjects if any of the type I or response-conditional type II hit or false alarm rates lie at the extremities ( $<.05$  or  $>.95$ ). Simulations found narrow exclusion criteria to lead to greater variance of meta- $d'$  but smaller bias than wide exclusion criteria. To determine which set of criteria will minimise distortion of data from any specific paradigm, we recommend using the Matlab code included in Barrett et al. (2013), which can simulate experiments and estimate the bias and variance in meta- $d'$  in the specific scenario.

In summary, both versions of meta- $d'$  invert the calculation of type II performance from type I performance into a calculation of estimated type I performance given type II performance. In this way many conceptual and theoretical problems relating to computing an overall measure of metacognition are avoided. Moreover, these problems also seem to be avoided in practice. Although there is, as yet, no single, optimal computation for meta- $d'$  it looks like meta- $d'_b$  is more robust to non-traditional SDT models whereas meta- $d'_{SSE}$  is less biased in small samples. The behaviour of meta- $d'$  computed by maximum likelihood estimation is as yet unexamined but is expected to be similar. The main drawbacks of the meta- $d'$  measures are that they are more noisy than the alternative measures discussed above, and that response-conditional versions do not improve on the stability of the alternative measures. Nevertheless, these measures are highly promising they are most promising for capturing metacognition independently of response biases. In summary, these measures will give stable and meaningful results when sufficient trials are obtained and the standard assumptions of SDT hold to reasonable approximation.

### **Measuring consciousness using type II signal detection theory and future extensions**

While we hope that we have given a thorough account of type II signal detection measures, how these relate to measuring consciousness is still a matter of debate. While metacognition may have been seen as outside of the reach of rigorous measurement, recent work hopefully renders this view unwarranted; we seem to now have an understanding of how properly to measure metacognition. Two questions now arise: (i) How can we use measures of metacognition to deepen our understanding of consciousness?; (ii) How can we extend SDT models to incorporate the range of cognitive processes we know to be modulators of consciousness?

To address the first question, there are arguments in the literature for using metacognition as a robust measure of visual awareness (Kunimoto & Miller 2001; Persuade et al 2007). These claim that confidence (or certain measures of confidence) taps in to the subjective states which underlie awareness. Moreover, it could be argued that representing a state is equivalent to being conscious

of that state (Higher-Order Thought theory), therefore if accuracy (the state) and confidence (the representation) correspond then the state must be consciously accessed. However, although in most cases it would be reasonable to assume that confidence would indeed correspond with accuracy for consciously but not unconsciously perceived stimuli, this presumption was violated in blindsight patient GY. GY demonstrated above chance metacognition (Evans & Azzopardi, 2007) yet is clearly unaware of visual stimuli in the blind field (Persaud et al. 2007). Therefore, while under certain circumstances we might (carefully) be able to use metacognition as a proxy measure of visual awareness or conscious knowledge, for a more rigorous assessment of unawareness we would hope to see a convergence with other measures which indicate unawareness – absence of EEG correlates such as the P300, for example. Metacognition does not wholly encapsulate all facets of consciousness; however given that it involves a complex translation of information and is a function of an experience (an experience of confidence) it clearly taps into the subjective and representational. It seems clear then, that metacognition remains an important concept in consciousness science even without its potential use as a direct index of awareness.

There is a debate to be had about how we should interpret a type II  $D'$  measure with relation to awareness. Imagine participants A and B take part in a psychophysical detection task. If A's meta- $d'$  is twice that of B, are they "twice as aware" of the stimulus? Are they twice as *often* aware or twice as *likely to be* aware of the stimulus? When metacognition is at chance it is much easier to interpret the results in relation to awareness than when making relative judgements between above-chance values. Moreover, if we consider response-conditional type II measures, their interpretation with relation to awareness is clearer for those trials reported as absent. This was described by Kanai et al. (2010) as an index of invisibility (SDI) - the extent to which subjective blindness is due to physical weakness (chance metacognition) of the stimulus rather than inattention (above chance metacognition). However how do we interpret metacognition for trials where the participant has reported presence? This comparison seems to tap the executive processes involved in decision making more, that is, metacognition directly. If this is the case, we might expect patient GY to demonstrate above chance overall metacognitive performance, but chance metacognition for reported absent trials.

In our second question we asked how signal detection theoretic models might be extended. It is well-known that top-down influences of attention (Sergent et al. 2013), expectation (Melloni, Schwiedrzik, Müller, Rodriguez, & Singer 2011), and emotion (Vuilleumier 2005) have effects on our conscious content. Therefore if we want to properly examine their effects on metacognition, on which, for some at least, there has been very little work, we need a way to incorporate these factors into type II SDT measures. Top-down attention has been modelled in the type I SDT framework by Rahnev et al. (2011) as a decrease in variance of the signal distribution. This follows from the claim that attention may reduce trial-by-trial signal variability. Similarly, it has been suggested that inattention may induce criterion jitter which, across trials, would increase the variance of the signal distribution relative to under attention (Ko & Lau 2012). With respect to the type II level, interestingly, it has recently been demonstrated that metacognitive performance for change detection is unaffected by inattention (Vandenbroucke et al. In press). Moreover, Sherman et al (under review) manipulated both attention and prior expectation and examined their effects on metacognition. Empirically, metacognition improved for expectation-congruent reports relative to expectation-incongruent reports, but only under full attention. They modelled the effects of high cognitive load as inducing criterion jitter at the type I level and found that, in line with empirical

results, this left confidence thresholds unaffected but reduced the dependence of metacognitive sensitivity on expectation. The effect of expectation was modelled by changing the type I signal and noise distributions to posterior probability distributions of target presence and target absence (respectively), given the expectation and the evidence. This shows us that SDT may indeed have the flexibility to adapt from a purely bottom-up framework to one in which all-important top down processes can be accounted for.

## Summary

In summary, we hope that we have convinced the reader of the value of signal detection theoretic models for measuring metacognitive performance. We have reviewed the literature on type II SDT measures and shown that there exist several measures which if used carefully and for an appropriate paradigm, now afford us the opportunity to assess metacognitive sensitivity robustly and rigorously. For the greatest range of robust measures (ROC curves, type II  $d'$ , meta- $d'$ ,  $r$ ), the paradigm should keep decision and confidence biases small and fixed, for example by limiting the number of each type of response a participant can make during the session. If biases are left free to vary, meta- $d'$  is the most robust, although it requires many trials in order to be stable. Moreover, meta- $d'$  is a useful measure for comparing type I sensitivity to type II (metacognitive) sensitivity since it is measured in the same units as type I  $d'$ .

Metacognition (in the manner that we discuss here) is still a relatively under-studied phenomenon outside of the field of memory and there are many questions that remain unanswered. Research into, for example, how different aspects of attention and expectation affect metacognition, how metacognitive performance on different tasks and between modalities are related, and how the neural substrates of metacognition arise in the brain, will contribute to our overall understanding of consciousness.

## References

- Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological methods*. doi:10.1037/a0033268
- Brown, G. S., & White, K. G. (2005). The optimal correction for estimating extreme discriminability. *Behavior Research Methods*, 37(3), 436–449.
- Clarke, F. R., Birdsall, T. G., & Tanner, W. P. (1959). Two types of ROC curves and definitions of parameters. *The Journal of the Acoustical Society of America*, 31(5), 629–630.
- Clifford, C. W. G., Arabzadeh, E., & Harris, J. a. (2008). Getting technical about awareness. *Trends in cognitive sciences*, 12(2), 54–8. doi:10.1016/j.tics.2007.11.009
- Dienes, Z., & Seth, A. K. (2010). Measuring any conscious content versus measuring the relevant conscious content : Comment on Sandberg et al . q. *Consciousness and Cognition*, 19(4), 1079–1080. doi:10.1016/j.concog.2010.03.009
- Dienes, Zoltán. (2008). Subjective measures of unconscious knowledge. *Progress in brain research*, 168(07), 49–64. doi:10.1016/S0079-6123(07)68005-4

- Dienes, Zoltán, & Scott, R. (2005). Measuring unconscious knowledge: distinguishing structural knowledge and judgment knowledge. *Psychological research*, 69(5-6), 338–51. doi:10.1007/s00426-004-0208-3
- Dienes, Zoltán, Scott, R. B., & Seth, A. K. (2010). Subjective measures of implicit knowledge that go beyond confidence: Reply to Overgaard et al. *Consciousness and Cognition*, 19(2), 685–686. doi:10.1016/j.concog.2010.01.010
- Evans, S., & Azzopardi, P. (2007). Evaluation of a “bias-free” measure of awareness. *Spatial vision*, 20(1-2), 61–77. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/17357716>
- Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychonomic bulletin & review*, 10(4), 843–76. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15000533>
- Gennaro, R. J. (2004). Higher-Order Theories of Consciousness : An Overview, (1974), 1–15.
- Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268), 732–764.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1.). New York: Wiley.
- Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of  $d'$ . *Behavior Research Methods, Instruments, & Computers*, 27(1), 46–51. doi:10.3758/BF03203619
- Kanai, R., Muggleton, N. G., & Walsh, V. (2008). TMS over the intraparietal sulcus induces perceptual fading. *Journal of neurophysiology*, 100(6), 3343–50. doi:10.1152/jn.90885.2008
- Kanai, R., Walsh, V., & Tseng, C. (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
- 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 of London. Series B, Biological sciences*, 367(1594), 1401–11. doi:10.1098/rstb.2011.0380
- Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General*, 124(3), 311–333. doi:10.1037//0096-3445.124.3.311
- Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, 10, 294–340.
- Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in cognitive sciences*, 15(8), 365–73. doi:10.1016/j.tics.2011.05.009
- Locke, J. (1700). *An essay concerning human understanding*.
- Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press.

- Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological bulletin*, *98*(1), 185–199.
- Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition*, *21*(1), 422–30. 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), 509–27. doi:10.1037/a0014876
- Mealor, A. D., & Dienes, Z. (2013). The speed of metacognition: taking time to get to know one's structural knowledge. *Consciousness and cognition*, *22*(1), 123–36. doi:10.1016/j.concog.2012.11.009
- Melloni, L., Schwiedrzik, C. M., Müller, N., Rodriguez, E., & Singer, W. (2011). Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. *Journal of Neuroscience*, *31*(4), 1386–1396. doi:10.1523/JNEUROSCI.4570-10.2011
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological bulletin*, *95*(1), 109–33. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/6544431>
- Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature neuroscience*, *10*(2), 257–61. doi:10.1038/nn1840
- Pollack, I. (1959). On Indices of Signal and Response Discriminability. *Journal of the Acoustical Society of America*, *31*, 1031.
- Rahnev, D., Maniscalco, B., Graves, T., Huang, E., De Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature neuroscience*, *14*(12), 1513–5. doi:10.1038/nn.2948
- Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies*, *49*(3), 329–359.
- 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
- Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness : Is one measure better than the other ? *Consciousness and Cognition*, *19*(4), 1069–1078. doi:10.1016/j.concog.2009.12.013
- Sergent, C., Wyart, V., Babo-Rebelo, M., Cohen, L., Naccache, L., & Tallon-Baudry, C. (2013). Cueing attention after the stimulus is gone can retrospectively trigger conscious perception. *Current biology : CB*, *23*(2), 150–5. doi:10.1016/j.cub.2012.11.047
- Seth, A. K. (2008). Post-decision wagering measures metacognitive content, not sensory consciousness. *Consciousness and cognition*, *17*(3), 981–3. doi:10.1016/j.concog.2007.05.008

- Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in cognitive sciences*, 12(8), 314–21. doi:10.1016/j.tics.2008.04.008
- Sherman, M. T., Seth, A. K., Barrett, A. B., & Kanai, R. (n.d.). Prior expectation facilitates metacognition for perceptual judgments.
- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of experimental psychology. General*, 117(1), 34–50. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/2966230>
- Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and cognition*, 20(4), 1787–92. doi:10.1016/j.concog.2010.12.011
- Vandenbroucke, A. R. E., Sligte, I. G., Barrett, A. B., Seth, A. K., Fahrenfort, J. J., & Lamme, V. A. F. (n.d.). Accurate metacognition for unattended visual representations.
- Vuilleumier, P. (2005). How brains beware : neural mechanisms of emotional attention, 9(12). doi:10.1016/j.tics.2005.10.011
- Wierzchoń, M., Asanowicz, D., Paulewicz, B., & Cleeremans, A. (2012). Subjective measures of consciousness in artificial grammar learning task. *Consciousness and cognition*, 21(3), 1141–53. doi:10.1016/j.concog.2012.05.012
- Wilimzig, C., & Fahle, M. (2008). Spatial attention increases performance but not subjective confidence in a discrimination task, 8, 1–10. doi:10.1167/8.5.7.Introduction

## Tables

Table 1. Type I SDT response variables

	Respond “present”	Respond “absent”
Signal present	Hit	Miss
Signal absent	False Alarm	Correct Rejection

Table 2. Type II d’ response variables

	Correct	Incorrect
Confident	Type II Hit	Type II False Alarm
Guess	Type II Miss	Type II Correct Rejection

Table 3. Response-conditional type II  $d'$

	Report present		Report absent	
	Correct (Hit)	Incorrect (False alarm)	Correct (Correct rejection)	Incorrect (Miss)
Confident	Type II Hit	Type II False Alarm	Type II Hit	Type II False Alarm
Guess	Type II Miss	Type II Correct Rejection	Type II Miss	Type II Correct Rejection

## Figures

Figure 1

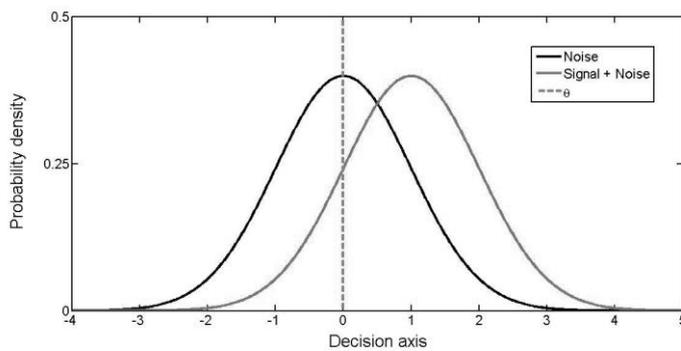
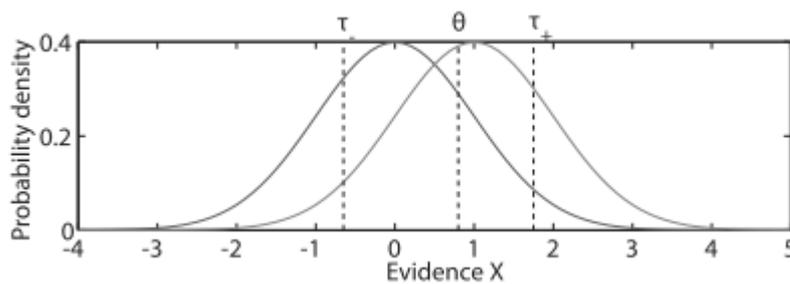


Figure 2



<sup>i</sup> These terms are distinct, however, from metacognitive awareness which is usually used to describe the phenomenal state. For example, feelings of familiarity with stimuli would indicate metacognitive awareness.