Chapter 6

Inferences about consciousness using subjective reports of confidence

Maxine T. Sherman, Adam B. Barrett, and Ryota Kanai

Introduction

An important aspect of consciousness is the ability to reflect upon one's own thoughts, an insight that 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 (Rosenthal 1986; Gennaro 2004; Lau and Rosenthal 2011), which posit that it is those states for which we have some representation or conceptualization 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 operationalize this ability as metacognitive sensitivity, performance, or accuracy, terms used interchangeably[1]. 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 participants are conscious of the stimulus, then they should be confident in their (correct) responses. In the case of stimulus detection this is most intuitive for trials on which a target is present rather than absent. The argument applies in the same way for measuring conscious knowledge; unconscious knowledge may generate feelings of knowing (Nelson 1984; Koriat 1995) or familiarity (Dienes and Scott 2005; Dienes et al. 2010). Thus we can use

[1] 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.

metacognitive performance as a proxy measure of awareness (Seth et al. 2008), although with caution (for debates in the literature see Kunimoto et al. 2001; Persaud et al. 2007; Seth 2008; Rounis et al. 2010; Song et al. 2011).

This chapter begins with a brief overview of type 1 signal detection theory (SDT), which is often used to calculate objective task performance. For a more thorough account of SDT we recommend Green and Sweets (1966) and Macmillan and Creelman (2004). We next cover ways in which researchers may measure confidence, and what would constitute a good measure of metacognition. We then move to a discussion of, largely, type 2 SDT measures of metacognition. All measures quantify metacognitive performance by examining the correspondence between the type 1 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 2 *d'*, type 2 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 1 SDT

SDT (Green and Swets 1966; Macmillan and 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 6.1. It is assumed that the probabilities of the stimulation being caused by noise and being caused by a noisy signal can each be modeled 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 modeled as a horizontal intercept called *c* or θ. In yes/no tasks, this threshold is often expressed as β, which represents the ratio of the likelihood of obtaining that signal strength on a signal trial to that on a noise trial. A β of 1 represents a bias-free observer, β greater than 1 represents a bias towards reporting noise, whereas β less than 1 represents a bias towards reporting a signal. Similarly, when considering decision threshold, 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."

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 these two distributions are equal. The second is that both the signal and the noise distributions are indeed Gaussian. The first of these

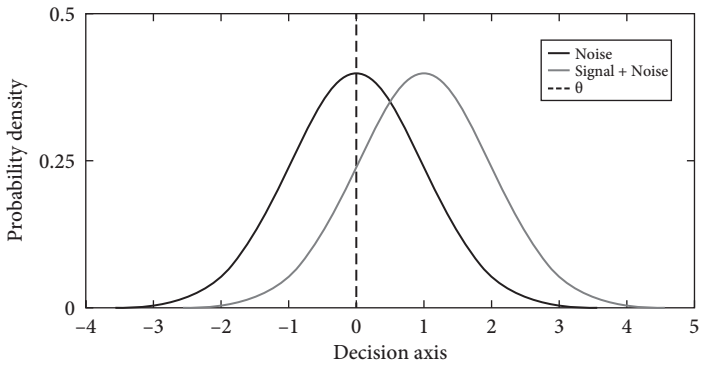

Fig. 6.1 Type 1 SDT. The black Gaussian is the probability density function for the evidence on a noise trial. The gray Gaussian is the probability density function for the evidence on a signal (plus noise) trial. These Gaussians are transformed so that the noise distribution is a standardized Gaussian (mean = 0, std. dev. = 1). The difference between their means (the peaks) is detection sensitivity d'. The dashed line θ represents the decision threshold. If the evidence takes a higher value than θ the response will be "signal" and if it is less the response will be "noise."

assumptions is sometimes problematic; 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 best fit by an unequal variances model.

In order to calculate d' the researcher collects data in a 2 × 2 design such that a signal is present or absent and the participant can be correct or incorrect. This leads to a variety of response variables as shown in Table 6.1. We can then calculate the following:

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

and

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

For hit rate h and false alarm rate f, task performance d' can then be calculated from $d' = \Phi^{-1}(h) - \Phi^{-1}(f)$ and the decision threshold θ can be calculated from $\theta = -\Phi^{-1}(f)$, where Φ^{-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, as in Fig. 6.1.

Table 6.1 Type 1 response categories.

	Respond "present"	**Respond "absent"**
Signal present	Hit	Miss
Signal absent	False alarm	Correct rejection

If one 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 1 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 if one can assume that *s* has been estimated to a good degree of accuracy (Macmillan and 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 Φ^{-1} function going to plus/minus infinity at 1/0. For *d'* to be finite, the hit and false alarm rates must 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 that may affect the decision threshold, 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 that are problematic. Here, in an experimental setup with *n* signal trials and (*N* – *n*) noise trials, a zero hit or false alarm rate would be replaced with 1/2*n* or 1/2(*N* – *n*), respectively. A hit or false alarm rate equal to 1 would be replaced with 1 – (1/2*n*) 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 1/2*N* rule (Macmillan and Kaplan 1985). An alternative transformation, the log-linear transformation proposed by Snodgrass and Corwin (1988), is to add 0.5 to all data cells (total hits, false alarms, correct rejections, and misses), regardless of whether they are problematic or not. This is advantageous in that all data are treated equally. Moreover, this correction can be considered a (Bayesian) unit prior, in which the prior hypothesis is that *d'* and θ are equal to zero (Barrett et al. 2013; Mealor and Dienes 2013). Hautus (1995) modeled the effects of both of these transformations on *d'* using Monte-Carlo simulated datasets. He found that both transformations can bias *d'* measures, and that while the log-linear rule systematically underestimated *d'*, the 1/2*N* rule was more biased and could distort *d'* in either direction. Therefore, although the log-linear rule is recommended over its counterpart, both in principal and practically, ideally data would

be collected such that the risk of obtaining troublesome data is minimized. An evaluation of numerous alternative transformations by Brown and White (2005) concluded the same as Hautus.

Type 1 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 that 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 when a response scale of length n has been used, there will be $n - 1$ ways to partition responses into hypothetical levels of decision threshold. 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 3–6 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 decision threshold θ (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:

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

A benefit of using SDT's *d'* over plotting an ROC curve (assuming the presence of both S_1 and S_2 trial types) 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 hypothesize a change in *h* but not *f*. Kanai et al. (2008) found that transcranial magnetic stimulation (TMS) 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 participants' judgment of their own accuracy one must compute both an accuracy measure and a confidence measure. Typically, experimental designs include some objective task such as target detection or word recall on which objective performance can be measured. To measure metacognitive sensitivity we use what is known as "the type 2 task," first coined by Clarke et al. (1959) and Pollack (1959), and so-called in reference to the aforementioned type 1 task of making decisions or judgments about the "state of the world." The type 2 task is to evaluate the accuracy of one's own decision. Galvin et al. (2003) discuss the type 2 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 1 decision does not make it a type 2 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 1 rating, no matter when it occurs.*"

Following this, it is advised that the confidence judgment 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 judgment; this suggests an asymmetry in the trustworthiness of the objective (type 1) and subjective (type 2) responses. If we instead take type 1 decisions as those that are about the state of the world, then we can take type 2 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 judgment 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 onto 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 on which participants have reported no confidence (i.e. we cannot infer this from low confidence). Therefore a rating scale should only be symmetrically collapsed onto 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 (PAS) could be used instead. This scale asks participants to rate the subjective visibility of their percept on a scale of 1 to 4 and is advisable for simple (e.g. stimulus detection) rather than complex (e.g. stimulus discrimination) designs (Dienes and Seth 2010; Sandberg et al. 2010), as the conscious content itself is not probed. This is discussed in more detail in Chapter 11.

Instead of requesting a type 1 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 and Fahle 2008). This, however, has not been verified for other type 2 measures. When using metacognition to assess the presence of conscious structural knowledge, one-step 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.

Measuring metacognition

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 TMS to prefrontal cortex (PFC), Rounis et al. (2010) were able to implicate this area in metacognitive sensitivity. They used bias-invariant (type 2) meta-*d*' (discussed in the section "Meta-*d*' and meta-*d*' balance") 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 2 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 judgment improves. This result was successfully computationally modeled using an SDT model with decision and confidence criteria modulated by prior expectations.

SDT 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 1 and type 2 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 behavior (for example, increased confidence for correct reports, known as type 2 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 1 response, i.e. to use a so-called response-conditional measure of metacognition. For example, in a target detection experiment, one has the classes "Report present" and "Report absent". Kanai et al. (2010) defined the subjective discriminability index (SDI) as a measure of subjective unawareness of stimuli, based on response-conditional type 2 ROC curves (see below). This can also be applied to type 2 *d'*. Specifically, by using only trials on which subjects reported absence of a target (type 1 correct rejections and misses) in the type 2 calculation, one measures 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 "Report 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 the 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 φ. These are equal in the binary case, but distinct for the non-binary case.

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

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

where S_x and S_y are the sample standard deviations of X and Y, respectively.

Alternatively, the phi correlation coefficient is calculated as

$$\varphi^2 = \frac{\chi^2}{n},$$

where χ^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 formula

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

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-SDT measure) is that r and φ 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 several of the proposed measures of metacognition.

Goodman-Kruskal gamma coefficient

The Goodman–Kruskal gamma coefficient (G) (Goodman and Kruskal 1954) is a non-parametric analogue of the SDT 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 × 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 was hoped to also be a flexible measure of metacognitive accuracy when applied to type 2 data (Nelson 1984). Task performance V is characterized as follows for a 2 × 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:

$$V = \frac{\sum hits \times \sum correct\ rejections}{\left(\sum hits \times \sum correct\ rejections\right) + \left(\sum misses \times \sum false\ alarms\right)}.$$

The gamma coefficient is then given by

$$G = 2V - 1 = \frac{\left(\sum hits \times \sum correct\ rejections\right) - \left(\sum misses \times \sum false\ alarms\right)}{\left(\sum hits \times \sum correct\ rejections\right) + \left(\sum misses \times \sum false\ alarms\right)}.$$

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 × 2) case, so for a thorough explanation we refer the reader to Masson and Rotello (2009).

In order to verify *G*'s supposed invariance to bias and distributional assumptions, Masson and Rotello (2009) simulated datasets in which metacognitive sensitivity was fixed and calculated *G*. More specifically, a 2AFC task was modeled 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. In fact, 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 2 *d'*

Type 2 SDT extends the logic of its type 1 counterpart by using confidence reports to map onto detection accuracy (Kunimoto et al. 2001; Macmillan and Creelman 2004). It assumes that correct and incorrect responses can be plotted on a type 2 decision axis as Gaussian random variables, analogously to the signal and noise distributions in type 1 SDT. The distance between the peaks of the distributions gives us our measure of metacognitive sensitivity, type 2 *d'*.

As shown in Tables 6.2 and 6.3, type 2 variables are computed analogously to type 1 variables, but instead of examining the correspondence between signal and response, response accuracy and confidence are compared. We define the type 2 hit as a confident and correct response, a type 2 false alarm as a confident but incorrect response, a type 2 miss as a correct but unconfident response, and a type 2 correct rejection as an appropriately unconfident, incorrect response. Metacognitive performance type 2 *d'* is then calculated

Table 6.2 Type 2 response categories.

	Correct	**Incorrect**
Confident	Type 2 hit	Type 2 false alarm
Guess	Type 2 miss	Type 2 correct rejection

Table 6.3 Response-conditional type 2 response categories.

	Report present		**Report absent**	
	Correct (Hit)	**Incorrect (False alarm)**	**Correct (Correct rejection)**	**Incorrect (Miss)**
Confident	Type 2 hit	Type 2 false alarm	Type 2 hit	Type 2 false alarm
Guess	Type 2 miss	Type 2 correct rejection	Type 2 miss	Type 2 correct rejection

analogously to type 1 *d'*—by subtracting the normalized type 2 false alarm rate from the normalized type 2 hit rate. The type 2 decision threshold θ 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 hold neither empirically (Evans and Azzopardi 2007) nor theoretically (Barrett et al. 2013) when type 1 and type 2 decisions are made based on the same evidence. Indeed, Barrett et al. (2013) found that under certain circumstances, type 2 *d'* is highly unstable. For example, if the type 1 criterion is placed where the noise and signal and noise distributions intersect, then type 2 *d'* is maximized when the observer is maximally unconfident, which would be a nonsensical and maladaptive 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 1 *d'*. Importantly, these analyses demonstrate a high reliance of type 2 *d'* on decision and confidence thresholds. The behavior of type 2 *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 2 *d'*. The assumption that type 1 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 2 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 and Lau 2012). Indeed, Galvin et al. (2003) have demonstrated that the type 1 decision axis (e.g. signal strength) cannot be transformed in such a way that the type 2 evidence distributions for correct and incorrect decisions are Gaussian. Thus if a single pathway underlies both type 1 and type 2 decisions then type 2 *d'* is not a measure of metacognition that arises naturally from SDT modeling.

Despite these problems, type 2 *d'* can still be useful as a basic measure of type 2 performance in some scenarios. Specifically, this is the case if response bias is small. Moreover, it is useful for situations in which 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 the section "Meta-*d'* and Meta-*d'* balance"), are too noisy to yield significant results.

Type 2 SDT

Computing type 2 *d'* is not the only way to apply SDT to measuring metacognition. Extension of the type 1 model into the type 2 domain can be achieved by overlaying confidence thresholds τ_+ and τ_- on the type 1 decision axis, as illustrated in Figure 6.2, such that confidence is high when $X < \tau_-$ or $X > \tau_+$ and low otherwise. Although this renders type 2 *d'* an unprincipled measure, it invites certain promising alternatives, namely type 2 ROC curves and meta-*d'*, as described in the section "Meta-*d'* and meta-*d'* balance".

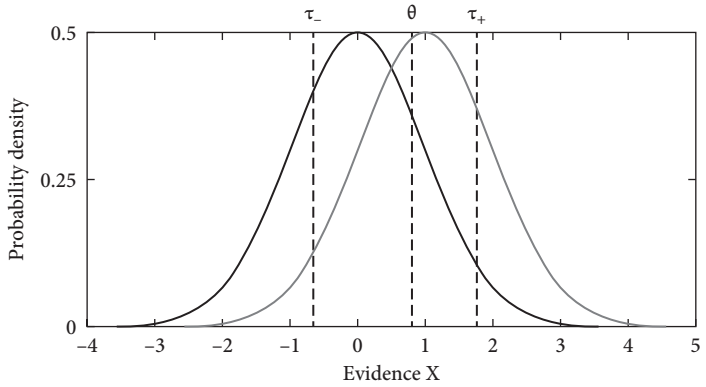

Fig. 6.2 Type 2 SDT. The Gaussian distributions represent the probability distributions for the evidence as in Fig. 6.1. Again θ represents the decision (type I) threshold, but here confidence thresholds τ_+ and τ_- are added. When the evidence lies between the two confidence thresholds a "guess" response is given. Otherwise, the response is "confident".

Type 2 ROC curves

While the type 1 ROC curve plots the probability of type 1 hits against the probability of type 1 false alarms for each level of criterion θ, the type 2 ROC curve plots the probability, for some fixed type 1 decision threshold θ, type 2 hit rate against type 2 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 2 level there are two thresholds, τ_+ and τ_- (confidence thresholds for positive and negative type 1 responses, respectively), for the response-unconditional case, three parameters are left to vary freely (θ, τ_+, and τ_-), rendering an infinite number of type 2 ROCs. Even if one fixes θ at the empirically observed level, the type 2 ROC is still not unique.

There are three potential solutions to this in the current literature. 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 θ and is straightforward to compute. However, the authors still found a strong dependence of the area under the ROC curve θ on θ.

An alternative measure, proposed by Clifford et al. (2008), suggested that one could compare the type 1 ROC curve based on a confidence rating scale with the ROC curve obtained by manipulating θ experimentally. That is, if one manipulates the physical properties of the stimulus such that response threshold changes (e.g. threshold contrast), then one can plot the false alarm rate/hit rate trade-off across artificially induced criterion shifts. This is the traditional type 1 ROC curve. We can compare this with an alternative type 1 ROC in which changes in criterion are modeled by differentially bisecting an n point rating scale into "signal" and "noise" responses 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 1 and the type 2

decision. Thus, Clifford and colleagues proposed their divergence as a measure of metacognition. Again, though, the degree of divergence is not in general independent of type 1 response bias.

Finally, Barrett et al. (2013) constructed the SDT-optimal type 2 ROC curve; the type 2 ROC curve that, for fixed θ and fixed type 2 false alarm rate (F), gives us the greatest type 2 hit rate (H), H_{max} (and therefore type 2 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 θ; 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 1 (decision) threshold is fixed, a unique type 2 ROC curve is obtained by varying a single confidence threshold. The response-negative area under the type 2 ROC curve forms the basis for Kanai and colleagues' (2010) SDI measure.

Meta-*d'* and meta-*d'* balance

Meta-*d'* (Maniscalco and Lau 2012) and meta-*d'* balance (written here as meta-d'_b; Barrett et al. 2013) are currently the gold standard in measures of metacognition. While type 2 *d'*, as mentioned in the section "Type 2 *d'*", computes metacognitive sensitivity as a function of accuracy and confidence, meta-*d'* computes the (type 1) accuracy that would be expected, given the type 2 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 2 *d'*, on the other hand, is formulated in different units from type 1 *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 1 information than the participant, rendering his or her performance suboptimal. A meta-*d'* of greater than *d'* is not possible on the standard SDT model (Fig. 6.2) but is possible if the observer has more information for making the type 2 decision than for making the type 1 decision, for example, after having had feedback on the type 1 decision or having had to make a speedy type 1 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 2 performance following a positive type 1 response and the other obtained by considering type 2 performance following a negative type 1 response. All existing approaches fix the type 1 response bias (the relative type 1 threshold β) 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 and Lau (2012) adopt a data-driven approach, by proposing two methods for finding the simultaneous best fit: minimizing the sum of the

squares of the errors (SSE) leads to meta-d'_{SSE}, while maximum likelihood estimation (MLE) 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, permits 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 1 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 and 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 and Lau (2012) write that if meta-d'_{SSE} is being used, it would be preferable to utilize a 2AFC task than a target detection task because target detection tasks are generally modeled 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 1 and type 2 levels, or when type 1 criterion is jittered across trials, representing fluctuations in attention), both versions remain reasonably robust, showing some variation when type 2 thresholds are varied, but little variation when the type 1 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 positive and negative response-conditional metacognitive performance. Under signal-degradation, signal-enhancement, and criterion jitter models, when variances are equal, both measures were largely invariant to changes in type 1 and type 2 thresholds.

Barrett et al. (2013) also looked at the behavior of both meta-*d*' measures on finite simulated datasets 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 1 or type 2 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 1 or response-conditional type 2 hit rates or false alarm rates are zero or one. These obviously maximize the number of subjects retained. An alternative choice is to use "wide exclusion criteria" which exclude subjects if any of the type 1 or response-conditional type 2 hit or false alarm rates lie at the extremities (< 0.05 or > 0.95). 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 minimize 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 2 performance from type 1 performance into a calculation of estimated type 1 performance given type 2 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 behavior of meta-*d*' computed by maximum likelihood estimation meta-d'_{MLE} 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 promising for reliably 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 2 SDT and future extensions

While we hope that we have given a thorough account of type 2 signal detection measures, how these relate to measuring consciousness is still a matter of debate. While metacognition may have been seen as outside 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: (1) How can we use measures of metacognition to deepen our understanding of consciousness? (2) 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 et al. 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 (HOT 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 and 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 that 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 the value of a metacognition measure with relation to awareness. Imagine that participants A and B take part in a psychophysical detection task. If A's meta-*d'* is twice that of B, is A "twice as aware" of the stimulus? Is A 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 judgments between above-chance values. Moreover, if we consider response-conditional type 2 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-level metacognition for reported absent trials.

In our second question we asked how SDT models might be extended. It is well known that top-down influences of attention (Sergent et al. 2013), expectation (Melloni et al. 2011), and emotion (Vuilleumier 2005) have effects on our conscious content. Therefore if we want to properly examine their effects on metacognition, we need a way to incorporate these factors into type 2 SDT models. Top-down attention has been modeled in the type 1 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 and Lau 2012). With respect to the type 2 level, interestingly, it has recently been demonstrated that metacognitive performance for change detection is unaffected by inattention (Vandenbroucke et al. 2014). 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 independently of attentional allocation. The effect of expectation was modeled by changing the type 1 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.

Conclusions

In summary, we hope that we have convinced the reader of the value of SDT models for measuring metacognitive performance. We have reviewed the literature on type 2 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 2 *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 1 sensitivity to type 2 (metacognitive) sensitivity since it is measured in the same units as type 1 *d'*.

Metacognition (in the manner that we discuss here) is still a relatively understudied 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., and Seth, A.K. (2013) Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, **18**(4), 535–552.

Brown, G.S. and White, K.G. (2005) The optimal correction for estimating extreme discriminability. *Behavior Research Methods*, **37**(3), 436–449.

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

Clifford, C.W.G., Arabzadeh, E., and Harris, J.A. (2008) Getting technical about awareness. *Trends in Cognitive Sciences*, **12**(2), 54–58.

Dienes, Z. (2008) Subjective measures of unconscious knowledge. *Progress in Brain Research*, **168**(07), 49–64.

Dienes, Z. and Scott, R. (2005) Measuring unconscious knowledge: distinguishing structural knowledge and judgment knowledge. *Psychological Research*, **69**(5–6), 338–351.

Dienes, Z. and Seth, A.K. (2010) Measuring any conscious content versus measuring the relevant conscious content: comment on Sandberg et al. *Consciousness and Cognition*, **19**(4), 1079–1080.

Dienes, Z., Scott, R.B., and 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.

Evans, S. and Azzopardi, P. (2007) Evaluation of a "bias-free" measure of awareness. *Spatial Vision*, **20**(1–2), 61–77.

Galvin, S.J., Podd, J.V, Drga, V., and Whitmore, J. (2003) Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychonomic Bulletin and Review*, **10**(4), 843–876.

Gennaro, R.J. (2004) *Higher-Order Theories of Consciousness: An Overview*, pp. 1–15. John Benjamins, Amsterdam.

Goodman, L.A. and Kruskal, W.H. (1954) Measures of association for cross classifications. *Journal of the American Statistical Association*, **49**(268), 732–764.

Green, D.M. and Swets, J.A. (1966) *Signal Detection Theory and Psychophysics* (Vol. 1). Wiley, New York.

Hautus, M.J. (1995) Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behavior Research Methods, Instruments, and Computers*, **27**(1), 46–51.

Kanai, R., Muggleton, N.G., and Walsh, V. (2008) TMS over the intraparietal sulcus induces perceptual fading. *Journal of Neurophysiology*, **100**(6), 3343–3350.

Kanai, R., Walsh, V., and Tseng, C. (2010) Subjective discriminability of invisibility: a framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and Cognition*, **19**(4), 1045–1057.

Ko, Y. and 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), 1401–1411.

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.

Kunimoto, C., Miller, J., and Pashler, H. (2001) Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, **10**, 294–340.

Lau, H. and Rosenthal, D. (2011) Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, **15**(8), 365–373.

Locke, J. (1700) *An Essay Concerning Human Understanding*. A. and J. Churchill, London.

Macmillan, N.A. and Creelman, C.D. (2004) *Detection Theory: A User's Guide*. Psychology Press, Hove.

Macmillan, N.A. and 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. and Lau, H. (2012) A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, **21**(1), 422–430.

Masson, M.E.J. and 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.

Mealor, A.D. and Dienes, Z. (2013) The speed of metacognition: taking time to get to know one's structural knowledge. *Consciousness and Cognition*, **22**(1), 123–136.

Melloni, L., Schwiedrzik, C.M., Müller, N., Rodriguez, E., and Singer, W. (2011) Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. *Journal of Neuroscience*, **31**(4), 1386–1396.

Nelson, T.O. (1984) A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, **95**(1), 109–133.

Persaud, N., McLeod, P., and Cowey, A. (2007) Post-decision wagering objectively measures awareness. *Nature Neuroscience*, **10**(2), 257–261.

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., and Lau, H. (2011) Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, **14**(12), 1513–1515.

Rosenthal, D.M. (1986) Two concepts of consciousness. *Philosophical Studies*, **49**(3), 329–359.

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

Sandberg, K., Timmermans, B., Overgaard, M., and Cleeremans, A. (2010) Measuring consciousness: is one measure better than the other? *Consciousness and Cognition*, **19**(4), 1069–1078.

Sergent, C., Wyart, V., Babo-Rebelo, M., Cohen, L., Naccache, L., and Tallon-Baudry, C. (2013) Cueing attention after the stimulus is gone can retrospectively trigger conscious perception. *Current Biology*, **23**(2), 150–155.

Seth, A.K. (2008) Post-decision wagering measures metacognitive content, not sensory consciousness. *Consciousness and Cognition*, **17**(3), 981–983.

Seth, A.K., Dienes, Z., Cleeremans, A., Overgaard, M., and Pessoa, L. (2008) Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences*, **12**(8), 314–321.

Sherman, M.T., Seth, A.K., Barrett, A.B., and Kanai, R. (under review) Prior expectation, but not attention, facilitates metacognition for perceptual decision.

Snodgrass, J.G. and Corwin, J. (1988) Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology. General*, **117**(1), 34–50.

Song, C., Kanai, R., Fleming, S.M., et al. (2011) Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition*, **20**(4), 1787–1792.

Vandenbroucke, A.R.E., Sligte, I.G., Barrett, A.B., Seth, A.K., Fahrenfort, J.J., and Lamme, V.A.F. (2014) Accurate metacognition for unattended visual representations. *Psychological Science*, **25**(4), 861–873.

Vuilleumier, P. (2005) How brains beware: neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, **9**(12).

Wierzchoń, M., Asanowicz, D., Paulewicz, B., and Cleeremans, A. (2012) Subjective measures of consciousness in artificial grammar learning task. *Consciousness and Cognition*, **21**(3), 1141–1153.

Wilimzig, C. and Fahle, M. (2008) Spatial attention increases performance but not subjective confidence in a discrimination task. *Journal of Vision*, **8**, 1–10.