# Confidence in consciousness research

Matthias Michel[1]

1. Center for Mind, Brain and Consciousness, New York University

**Author's version. Forthcoming in the *WIREs Cognitive Science*. Please cite the published version.**

**Abstract:** To study (un)conscious perception and test hypotheses about consciousness, researchers need procedures for determining whether subjects consciously perceive stimuli or not. This article is an introduction to a family of procedures called 'confidence-based procedures', which consist in interpreting metacognitive indicators as indicators of consciousness. I assess the validity and accuracy of these procedures, and answer a series of common objections to their use in consciousness research. I conclude that confidence-based procedures are valid for assessing consciousness, and, in most cases, accurate enough for our practical and scientific purposes.

To study (un)conscious perception and test hypotheses about consciousness, scientists need procedures for determining whether subjects consciously perceive stimuli or not—consciousness detection procedures (Irvine, 2012a, 2012b; Michel, 2021; Spener, 2020). This article discusses a family of detection procedures called 'confidence-based procedures' (Morales & Lau, forthcoming; Norman & Price, 2015).

Let's start with an example. Suppose you're in a consciousness science experiment. Masked stimuli briefly appear on a screen. Your task is to identify them. After each decision, you provide confidence ratings—report how confident you are that your decision was correct. After repeating this a thousand times or more, scientists analyze your data. They usually get two scores: your identification performance and your metacognitive performance—how good you were at evaluating your decisions. Confidence-based detection procedures consist in interpreting this metacognitive score as an indicator of consciousness. In other words, scientists assess whether you perceived stimuli consciously or unconsciously by analyzing how good you were at evaluating decisions about the visual features of the stimuli.

Many have found this way of assessing consciousness somewhat puzzling (e.g., Abid, 2018; Irvine, 2012a; Rosenthal, 2019). And who can blame them: aside from the vague intuition that unconscious decisions should be made with low confidence, little has been done to justify using confidence ratings in consciousness research (but see Morales & Lau, forthcoming; Norman & Price, 2015).

I try my best to justify confidence-based procedures in this article. Section 1 introduces the main detection procedures used in consciousness research as well as the general framework of Signal Detection Theory. Section 2 is a non-technical introduction to the main metacognitive indicators used in confidence-based procedures. Section 3 links these indicators to consciousness and evaluates whether one can use them to accurately assess consciousness. Section 4 shows why confidence-based procedures are better than alternative procedures. Section 5 answers eleven objections.

## 1. Detecting Consciousness, Perception, and Sensory Registration

### 1.1. Detecting Consciousness

Studying consciousness often requires answering two questions: (1) Do participants *perceive* the stimuli?[1] (2) Do they perceive them *consciously*? We will focus on the latter here. There are three main options for answering it: visibility-based procedures, confidence-based procedures, and 'objective' procedures (see Irvine (2013) and Spener (2020) for detailed introductions).

Visibility-based procedures are rather straightforward. Just ask participants whether they saw the stimulus or not in each trial. Scientists have used a variety of visibility scales to do that, from simple seen/not seen scales (Sergent & Dehaene, 2004), to the 'Perceptual Awareness Scale' (Ramsoy & Overgaard, 2004)—a scale mainly developed to encourage participants to report weak experiences. If a participant performs above chance on a task that requires seeing the stimulus—such as identifying a certain visual feature, but reports not seeing it, that's an indication that the participant saw that feature unconsciously.

Visibility-based procedures seem like an obvious choice. But the oldest attempts at investigating unconscious perception instead relied on confidence-based procedures (Peirce & Jastrow, 1884). I describe confidence-based procedures in detail in Sections 2 and 3. For now, here is the basic idea: if a participant can identify a stimulus above chance, but her confidence ratings indicate that she has no idea that she's able to do so, that's an indication that she wasn't conscious of the visual features used to successfully perform the task.

Finally, one could be skeptical of the very project of identifying instances of unconscious perception (Holender, 1986; Phillips, 2016, 2018, 2021; Reingold & Merikle, 1990; Snodgrass et al., 2004)—thereby collapsing questions (1) and (2) above. Objective procedures consist in interpreting indicators of perception as indicators of *conscious* perception. This is probably what we do in ordinary life: if you behave in ways indicating that you see a stimulus, I'll infer that you're conscious of it. I don't have to *ask* whether you *consciously* saw it in order to settle the consciousness question. Perception is 'conscious until proven otherwise' (Balsdon & Clifford, 2018).

[1] Depending on the purpose of the experiment, one might replace this question with more or less ambitious questions. For instance, researchers investigating unconscious working memory could ask: (1) Do participants encode stimuli in working memory? (2) Do they perceive the stimuli encoded in working memory consciously? (e.g. King et al. 2016; Trübutschek et al. 2019) Similarly, researchers investigating whether or not attention is sufficient for consciousness could ask: (1) Do participants attend to the stimuli? (2) Do they perceive the stimuli they attend to consciously? (e.g. Kentridge et al. 2008) In each case, one needs two indicators: an indicator of whatever capacity one is currently investigating, and an indicator of consciousness.

For the purpose of this article, I leave objective procedures aside and focus on the methods most commonly used in consciousness research: visibility-based and confidence-based procedures (LeDoux et al. 2020)[2].

Our goal is to understand how one can use confidence-based procedures to assess whether participants perceive stimuli *consciously* or *unconsciously*. To do so, we need to understand how researchers determine whether or not participants *perceive* stimuli in the first place. This question is answered with the tools of Signal Detection Theory (SDT) (Green & Swets, 1966; Macmillan & Creelman, 2005), and answering it requires drawing a distinction between perception and sensory registration.

## 1.2. Sensitivity, perception, and sensory registration

Suppose that your task is to detect a visual stimulus. Transducers transform light incoming from that stimulus into an increase in sensory activity in the relevant perceptual channel. This increase in sensory activity is a *signal* that the stimulus is present. Sensory signals like this one do not occur in a vacuum. Instead, the signal adds to the sensory activity already present in the relevant channel—spontaneous noise. This pre-existing activity can take various values, sometimes weaker, sometimes stronger, assumed to be normally distributed[3] (Figure 1). Adding a constant signal—an increase in sensory activity—to that noise shifts this distribution by an amount equal to the strength of the signal, which depends on the subject's *sensitivity* to the presence of the relevant feature[4]. If a subject is very sensitive to a feature, presentations of that feature lead to strong increases in sensory activity—strong signals—relative to the noise alone. One can thus quantify a subject's sensitivity to a feature (noted *d'*) by measuring the number of standard

[2] I take it that there is good evidence indicating that perception and consciousness sometimes dissociate, leading to instances of unconscious perception. This is the case, for instance, in the phenomenon of blindsight (Weiskrantz (2009), although see Phillips (2021), and Michel & Lau (2021) for a response).

[3] If noise is spontaneous random sensory activity, one can justify a Gaussian noise distribution by the central limit theorem (Thompson & Singh, 1967). But other assumptions can be adopted and motivated. For instance, Luce's choice axiom justifies logistic distributions in Luce's Choice Theory (1959), which is an alternative to SDT.

[4] Note the assumption that the signal adds a constant value to the noise—sensitivity to the stimulus does not vary across trials. While often assumed for simplicity (Peterson et al. 1954), this assumption is likely incorrect in naturalistic settings (Swets, 1986a, 1986b). In practice, varying sensitivity across trials leads to noise and signal distributions with unequal variance (DeCarlo, 2010).

deviations (z-score) between the mean amount of sensory activity corresponding to the noise alone and the mean amount of sensory activity generated by the noise and the signal.

A challenge for the subject (and her perceptual system) is that the same amount of sensory activity could either reflect meaningless noise or the presence of a stimulus 'out there'. The signal and noise distributions overlap. This challenge is solved by setting a *decision criterion*: a level of sensory activity above which the system outputs the decision that the stimulus is present. The optimal setting to avoid miss and false alarms is to set the criterion at the level of sensory activity at which the noise and signal distributions intersect. Any other criterion leads to *biased* decisions. Either the criterion is overly *liberal* on the amount of sensory activity that should count as indicating stimulus presence (high false alarm rate), or it is overly *conservative* (high miss rate). The beauty of SDT is that sensitivity (indicated by *d'*) and decision criterion values can both be estimated independently based on the participant's rates of hits, false alarms, miss, and correct rejections (Figure 1).

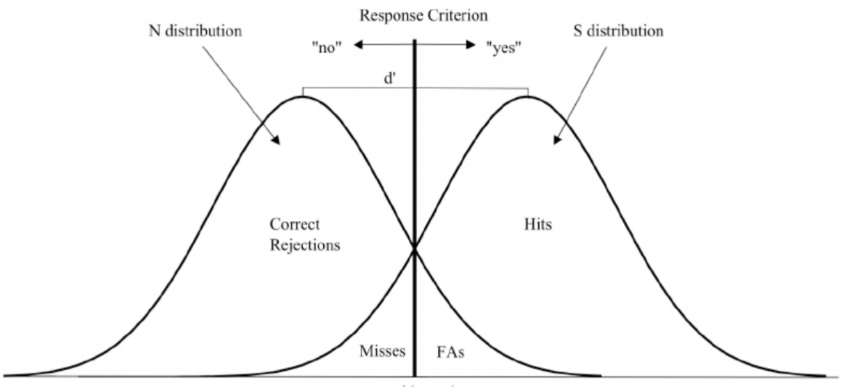


**Figure 1.** Source: Macmillan and Creelman (2005). Standard model of signal detection theory.

Whether a subject perceives a visual feature or not is often estimated with *d'*. Doing so amounts to interpreting the subject's sensitivity to the presence of a visual feature as an indicator of *perception* of that feature. At this point, it is important to recognize that *d'* is a measure of the sensitivity *of the subject* to a feature, not just the 'sensitivity' of her visual system. Mechanisms in my visual system might be 'sensitive' to features that *I* am not sensitive to, as a subject. This is crucial for understanding the distinction between *perception* and *sensory registration*. Let me illustrate.

A large fraction of neurons in visual area V1 are not only sensitive to the presence of stimuli in different locations of visual space, but also to the eye of origin of the visual input (Hubel & Wiesel, 1977). This eye-of-origin information is completely lost at the perceptual level:

subjects cannot discriminate the eye of origin of visual features even after weeks of training (Zhaoping & Xiao, 2016). Now, imagine what would happen if V1 were a small homunculus. This V1-homunculus would be sensitive to the difference between stimuli presented to the left and right eye. From its 'perspective', these stimuli would 'look' different. But from *your* perspective, they're indistinguishable. While neurons in V1 are sensitive to the eye of origin, *you*—as a subject—are not. Of course, since V1 is not a homunculus, we can't speak of V1, or anyone else for that matter, *perceiving* the eye-of-origin (whatever that means). Instead, we say that V1 *registers* information pertaining to the eye of origin.

Sensory registration of this kind can have a variety of downstream effects. For instance, since V1 neurons are sensitive to the eye of origin of the input, and since V1 encodes a visual salience map mediating exogenous attentional capture, attentional capture can be triggered by visual information subjects are not sensitive to (Zhaoping, 2008, 2019) (Figure 2). Other examples include, for instance, priming effects (Schmidt et al. 2010; Vorberg et al. 2003), attentional capture by masked stimuli (Hsieh et al. 2011; Zhang et al. 2012), or visual adaptation to flickers below the flicker fusion threshold (Shady et al. 2004; Vul & MacLeod, 2006).

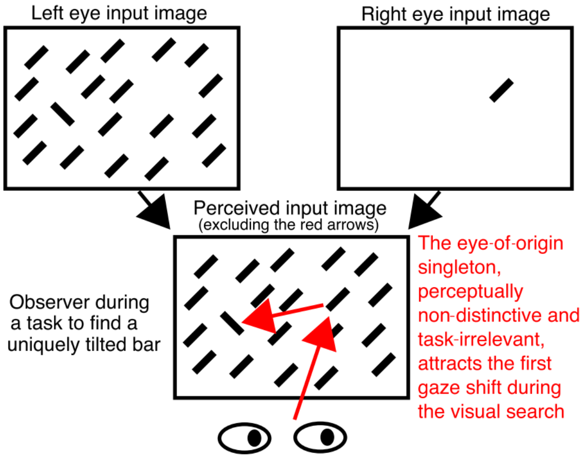


**Figure 2.** Source: Zhaoping (2019). V1 neuronal responses are sensitive to the eye-of-origin. Because of this, a singleton displayed in the right eye leads to a strong neural response at the corresponding location, even if it is perceptually non-distinctive. This sensory registration is used to compute a visual salience map, and thus automatically drives exogenous attentional capture to a *perceptually* non-distinctive and task-irrelevant location. This is an example of a downstream effect of unconscious sensory registration.

Effects of this kind do not necessarily indicate that the subject *perceived* the relevant stimuli. They just indicate that the subject's perceptual system *registered* the relevant information. Demonstrating unconscious sensory registration is relatively easy: one simply needs to find instances where the sensory system registers the presence of a visual feature—as demonstrated, for instance, by neural responses, or priming effects—without subject-level sensitivity (as indicated by *d'* = 0). Since one cannot consciously perceive a feature if one does not perceive that feature in the first place, sensory registration without perception has to occur *non-consciously*. So, demonstrating sensory registration of a feature when *d'* = 0 for detecting it amounts to demonstrating unconscious sensory registration[5]. Meanwhile, *d'* > 0 indicates that the *subject* is sensitive to some visual feature, not just her visual system. And I am going to assume that when a subject is sensitive to a feature, she perceives that feature[6].

We just saw how to assess perception with SDT, and how perception differs from sensory registration. We also saw how to find evidence of unconscious *sensory registration*. Now the question is: how can we demonstrate unconscious *perception*?

Unconscious perception is sensitivity to a visual feature or to the difference between two visual features in conditions in which the subject *feels* like she does not perceive that feature, or *feels* like the two features are indistinguishable. In other words, unconscious perception is sensitivity to a feature that feels just like no sensitivity to that feature—perceiving a feature unconsciously feels just like not perceiving it.

So, we need an indicator that we could interpret as indicating whether perceiving a given feature *feels like something for the subject*. Finding such an indicator has always been a central problem in consciousness research (Ledoux et al. 2020; Irvine, 2012a; Michel, 2020). I will now argue that *metacognitive* indicators are currently the best indicators of consciousness, even though they're far from perfect.

A final bit of terminological housekeeping before moving on. I will sometimes talk about sensory activity being 'conscious' or 'unconscious'. 'Conscious sensory activity' is sensory activity

[5] 'Chance-level' performance does not always indicate no perception. A variety of factors could explain bad performance on a visual task, like failures of memory, or failures to understand and comply with task instructions. *d'* = 0 is good evidence of *no perception* when these other factors are ruled out, which can be the case if visual tasks are well-designed.

[6] Here is a weak argument to support this claim: (1) Successful performance in a visual task (indicated by *d'*) is evidence that the participant is in a state allowing her to perform a voluntary action in accordance with some task instructions stored in memory; (2) Availability for voluntary action and integration with task instructions stored in memory indicate perception—presumably because they're evidence that the relevant representations are available for a 'central coordinating agency' (Burge, 2010; Shepherd & Mylopoulos, 2021; Quilty-Dunn, 2019). Therefore, successful performance in a visual task indicates perception. Premise (1) seems correct. But since there's no uncontroversial account of the personal/subpersonal distinction (Drayson, 2012; Taylor, 2020), premise (2) remains controversial.

in virtue of which a subject has a conscious experience as of perceiving what that activity signals. When sensory activity signals that there is a red dot out there, this sensory activity is 'conscious' if I consciously experience a red dot in virtue of the occurrence of that sensory activity. Unconscious sensory activity is sensory activity that occurs without a conscious experience of what that activity signals. Coming back to our example, when sensory activity in V1 signals that a visual feature is presented to the left eye, that sensory activity occurs without my consciously experiencing that feature *as being presented to the left eye*. So, sensory activity signaling the eye of origin is unconscious.

## 2. Metacognitive Indicators

### 2.1. Confidence ratings: The Naive View

Suppose that you participate in the experiment depicted in Figure 3. You have two tasks in each trial. Identify stimulus orientation. And report your confidence in your decision. The former is called a *Type-1 task*. The latter is a *Type-2 task*—it involves "discriminating between one's own correct and incorrect decisions" (Galvin et al. 2003, p.843; Clarke et al. 1959). Providing confidence ratings about perceptual decisions is a way of doing this. Since perceptual decisions are mental events, providing confidence ratings about your own perceptual decisions requires *metacognition*: the capacity to think about your own mental states.

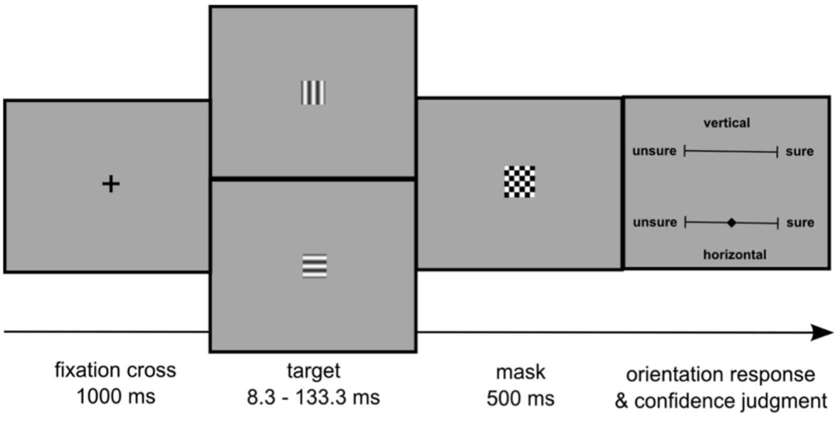


Figure 3. Source: Rausch et al. (2018). Subjects decide whether a masked stimulus is oriented vertically or horizontally (Type-1 task) and evaluate their decision (Type-2 task).

Subjects can provide confidence ratings on a variety of scales, such as binary scales, discrete scales with several confidence levels, or continuous percentage scales. The lowest confidence level is usually described as a 'complete guess': the response is as good as flipping a coin. Confidence then increases ordinally from there to the highest level[7]. Only two cases of confidence-based procedures do not rely on this kind of confidence ratings: post-decision wagering, where confidence ratings are replaced by bets on the accuracy of one's decisions (Persaud et al. 2007); and the confidence forced-choice task, where observers choose which of two perceptual decisions is most likely correct (Barthelmé & Mamassian, 2009, 2010; Peters & Lau, 2015; Mamassian, 2020).

Confidence ratings constitute raw data that scientists have to analyze. One way of interpreting them would be to assume a one-to-one mapping between confidence and consciousness of the stimulus. High confidence means conscious. Low confidence means not conscious. That's what I call the 'naive view'—a rather disastrous way of interpreting confidence ratings, or even visibility ratings for that matter.

There are two main reasons for rejecting the naive view. First, fluke responses. Subjects sometimes press the 'guess' button when they meant to press 'confident', or press 'guess' because they previously pressed 'confident' six times in a row, or press a random button to move on to the next trial (who will notice anyway?), or press 'guess' just because they've been instructed to use the entire scale, and so on. Because of the possibility of these 'fluke responses', any individual report shouldn't be interpreted too literally (Michel 2021).

Second, subjects vary in metacognitive bias: "the tendency to give high confidence ratings, all else being equal" (Fleming & Lau, 2014). Some people are highly self-confident.

[7] Since confidence scales are not interval scales, using statistical methods such as computing mean confidence, or an average increase in confidence is meaningless, strictly speaking (Stevens, 1946). 'Average confidence' only makes sense if one assumes that a subject's switch from report category 1 to 2 and her switch from report category 3 to 4 reflect the same increase in confidence. Experimenters often use parametric statistical analyses (e.g., t-test, ANOVAs, etc.) on confidence rating data, which, rigorously, only make sense for data represented on interval and ratio scales. Nevertheless, the prescriptive aspect of measurement scales for statistical analysis is the subject of a longstanding debate between measurement theorists and statisticians (e.g., Stevens, 1951; Lord, 1953; Gaito, 1980; Michell, 1986). We should distinguish the conceptual aspect of measurement scales from aspects that make a real practical difference. From a conceptual perspective, using certain types of statistical analyses on data that are not represented on ratio or interval scales is meaningless. But from a practical perspective using what could seem to be inappropriate methods might lead to results that are not much different from those one would obtain with more conceptually sound methods. It seems that Stevens considered this separation between the conceptual and practical aspects acceptable: "As a matter of fact, most of the scales used widely and effectively by psychologists are ordinal scales. In the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these scales … On the other hand, for this 'illegal' statisticizing there can be invoked a kind of pragmatic sanction: in numerous instances it leads to fruitful results" (Stevens, 1946, p.679).

They'll tend to have a *liberal* bias. Others aren't. They'll have a *conservative* bias. The problem is that bias and conscious perception can vary independently. To illustrate, suppose experimenters pay me $100 dollars each time I answer 'guess'. In these conditions, I'll answer 'guess' all the time—I'm only human after all. Accepting the naive view would lead one to conclude that unconscious perception can be bought. But my perceptual system doesn't take bribes. So, again, 'guess' or 'confident' reports cannot be interpreted too literally[8].

Rejecting the naive view requires giving up on the idea that we can determine whether a participant was conscious of a stimulus on a given trial. As argued by Fleming and Lau (2014):

> if one only has access to a single rating of performance, it is not possible to tease apart bias from sensitivity, nor measure efficiency. (...) In contrast, by collecting trial-by-trial measures of performance and metacognitive judgments we can build up a picture of an individual's bias, sensitivity and efficiency in a particular domain. (p.5)

Instead, one can assess consciousness based on *patterns of responses* obtained *throughout* an experiment or condition of an experiment, sometimes *across* subjects (Michel, 2021). As Timmermans and Cleeremans (2015) write, by doing so "one abandons the ability to establish, for any single stimulus, whether it was consciously perceived or not, simply because computing correlations requires many trials" (p.38).

## 2.2. Metacognitive sensitivity, bias, and efficiency

Let me now introduce metacognitive indicators, starting with metacognitive bias, sensitivity, and efficiency (Fleming & Lau, 2014). As I will show in Section 3, these indicators are crucial for assessing consciousness with confidence-based procedures. Metacognitive bias is a decrease or increase in confidence when performance is constant. Metacognitive sensitivity is the capacity to distinguish correct from incorrect decisions. Finally, metacognitive efficiency is the ability to distinguish correct from incorrect decisions, *given one's Type-1 performance* (Maniscalco & Lau,

[8] The naive view looms behind an unfortunate practice in consciousness research: binarizing data into 'low confidence' and 'high confidence' (or not seen/seen) categories and then analyzing 'low confidence' trials as if they were 'unconscious trials'. Doing so essentially amounts to selecting a sample of fluke responses and responses that perhaps come mostly from conservative observers, and then analyzing them as if they reflected unconscious perception (Schmidt, 2015; Shanks, 2017). Schmidt (2015) pointed out a sampling fallacy: suppose that subjects report an invariant stimulus as visible (or report high confidence) 95% of the time. By any standard, one should consider that the stimulus was clearly visible for the participants. Yet, as noted by Schmidt (2015), "the [not-seen-judgments-only] procedure would still invite us to analyze the few [not seen] trials for effects of unconscious perception" (p.36).

2012). This latter notion will be important for assessing consciousness (Section 3), but it's slightly more difficult to understand, so let me use an analogy.

Suppose that I never lost a single race—an impressive record. Does that make me a good runner? Not if I only accept to race against toddlers. My record depends on my competitors. But irrespective of my competitors, I am presumably just as good (or bad) a runner whether I'm competing against toddlers or Usain Bolt. Like my winning record, metacognitive sensitivity depends on task difficulty—as we'll see next (Galvin et al. 2003). But one might want to evaluate metacognition irrespective of task performance. For instance, one could aim to assess metacognitive performance across conditions where task performance differs. Measuring metacognitive efficiency allows researchers to do precisely this.

To further explain how to measure these constructs, starting with metacognitive sensitivity[9], let me introduce another concept from SDT: the Receiver Operating Characteristics (ROC) curve, or isosensitivity curve (Swets, 1973). Two subjects can have the same *d'* with different hit and false alarm rates (Figure 4). An ROC curve is the set of possible hit/false alarm rate pairs an observer can produce with a given *d'*, while the observer's bias determines which of those possible pairs she *actually* realizes. Since the shape of the ROC curve is independent from the observer's criterion setting, the area under the ROC curve (AUROC) measures the observer's sensitivity to a feature independently of her criterion setting.

[9] An initially appealing way to estimate Type-2 sensitivity would be to estimate it just as Type-1 *d'*, as suggested by Kunimoto et al. (2001). But *d'* is not 'contaminated' by response bias only because one assumes normal signal and noise distributions with equal variance (Macmillan & Creelman, 2005). To use the same method to compute Type 2 sensitivity, Kunimoto et al. (2001) had to assume the existence of Gaussian distributions of correct versus incorrect "confidence signals". But this is wrong, as demonstrated both through simulations as well as experimental evidence (Evans and Azzopardi, 2007; Galvin et al., 2003).

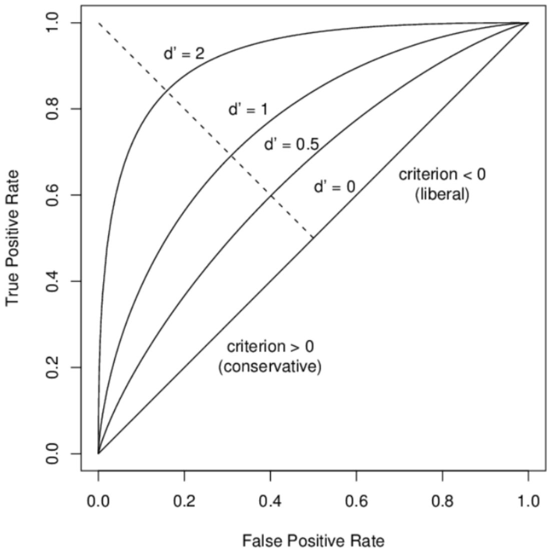
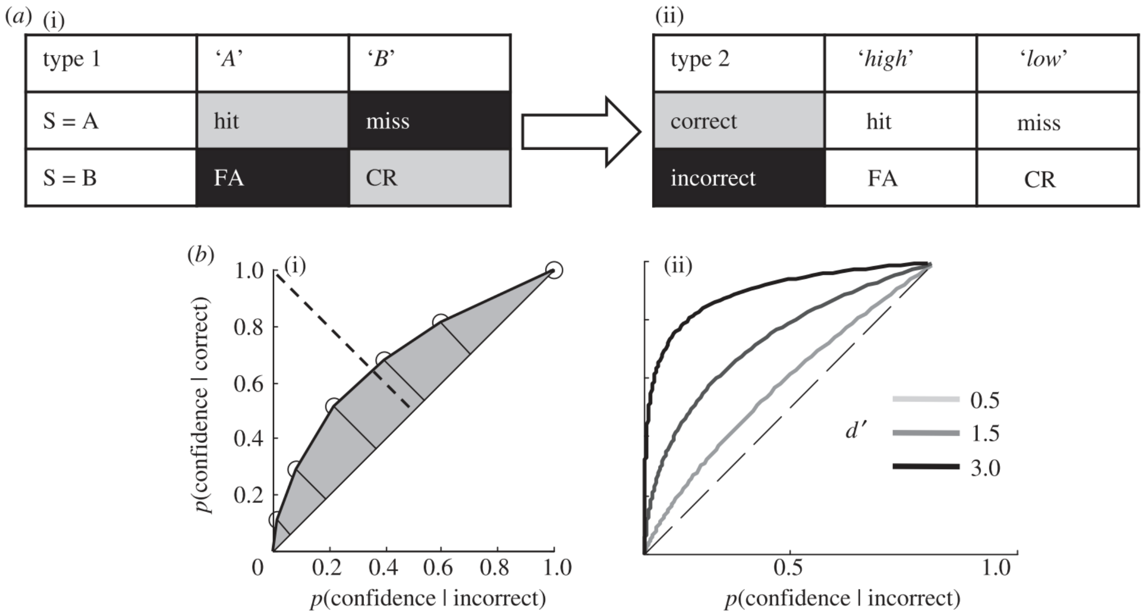


**Figure 4.** Adapted from Macmillan and Creelman (2005). ROC curves connecting locations with constant *d'*. The major diagonal is called the "chance line", since the hits and false alarm rates are equal, meaning that the subject is performing at chance level. Accuracy is perfect when the rate of hits (true positive rate) is equal to 1, and the rate of false alarms (false positive rate) equal to 0. A conservative criterion decreases the rate of hits and false alarms, and a liberal criterion increases the rate of hits and false alarms. Using the area under the ROC curve, one can assess sensitivity independently of the response criterion.

One can similarly estimate metacognitive sensitivity by defining Type-2 hits, false alarms, misses and correct rejections, and then computing a Type-2 equivalent of AUROC called AUROC2 (Fleming & Lau, 2014; Figure 5). Just as AUROC measures perceptual sensitivity, free from Type-1 biases, AUROC2 measures metacognitive sensitivity, free from Type-2 biases: it represents the observer's metacognitive sensitivity independently of her particular criterion setting on the Type-2 task (but see Shekhar & Rahnev, 2021; Xue et al. 2021).

AUROC2 indicates metacognitive *sensitivity*, not metacognitive *efficiency*, because it depends on sensitivity in the Type-1 task (Figure 5b). Suppose that a subject has no sensitivity to a feature at all. Metacognitive sensitivity should be equal to zero. But if stimuli always elicit a very high level of sensory activity compared to the noise alone (high signal strength), participants are more likely to be correct *and* more confident in their responses, thus driving metacognitive sensitivity up. As noted above, one might want to evaluate a participant's metacognition independently of her Type-1 performance—that's metacognitive *efficiency*.



(a) (i)

| type 1 | ‘*A*’ | ‘*B*’ |
|---|---|---|
| S = A | hit | miss |
| S = B | FA | CR |

(ii)

| type 2 | ‘*high*’ | ‘*low*’ |
|---|---|---|
| correct | hit | miss |
| incorrect | FA | CR |

**Figure 5.** Source: Fleming & Dolan (2012). (a) (i) Contingency table for Type-1 responses; (ii) Contingency table for Type-2 responses. (b) (i) Example of a type 2 ROC curve. Each point plots the type 2 false alarm rate on the x-axis against the type 2 hit rate on the y-axis for a given choice of a confidence criterion (H). The shaded area is the area under the curve (AUROC2), and represents metacognitive sensitivity. (ii) Predicted metacognitive sensitivity as a function of perceptual sensitivity. AUROC2 increases as *d’* increases.

The gold standard for measuring metacognitive efficiency is the M-ratio, itself based on an indicator called *meta-d’* (Maniscalco and Lau, 2012). The M-ratio is crucial for confidence-based procedures in consciousness research. Let me introduce the main idea with a thought experiment, vaguely inspired from Feigl’s (1958) auto-cerebroscope thought experiment.

Your task is to detect stimuli and rate your confidence in your perceptual decisions. After each perceptual decision, and before your confidence judgment, an advanced technology called an ‘SDT-scope’ displays the sensory activity elicited in your own perceptual system in the relevant perceptual channel during the trial. On this SDT-scope, sensory activity is displayed SDT-style: you can determine after each decision whether sensory activity was more likely drawn from the noise or signal distribution (Figure 1).

Experimenters tell you that you should use information displayed on the SDT-scope to rate your confidence in your perceptual decisions, such that your confidence ratings exactly match your probability of being correct. How should you use it?

You know that when sensory activity falls where the noise and signal distributions overlap, the probability of giving the correct response is close to chance. So you should answer 'guess'. But as sensory activity falls farther and farther from the criterion, the probability of giving the correct answer increases. So you can be more confident that you got it right. As Mamassian (2020) writes: "confidence evidence can be taken as the distance of the sensory sample away from the criterion" (p.621).

We can draw two lessons from this story. The first is that the probability of being correct scales with the strength of the sensory activity (relative to noise) on any given trial, being maximal for extreme values. The second is that an *ideal metacognitive observer* who would have access to the relevant information—like the information displayed on the SDT-scope—would be able to provide confidence ratings that precisely track the objective probability of being correct. In other words, we can know, for any Type-1 performance, what the *ideal* metacognitive performance should be. This is leveraged to compute the *meta-d'* indicator.

Using Bayesian ideal observer analysis, one can model an optimal (e.g., error-minimizing) metacognitive performance for any Type-1 performance (Galvin et al. 2003; Maniscalco & Lau, 2012, 2014). Which means that one can determine the metacognitive performance that an observer *would have had*, had she been an ideal metacognitive observer, given her *actual* Type-1 performance (*d'*). The reverse is also true. Based on a participant's *actual* metacognitive sensitivity, one can determine the Type-1 performance she would have had, had she been an ideal metacognitive observer. *Meta-d'* aims to measure just that: the performance that a subject would have had on the Type-1 task, given her *actual* Type-2 sensitivity, had she been an ideal metacognitive observer.

'Why should we care about *ideal* performance?', you ask. Here's an analogy. You transfer some water from one bucket to another. Since you're reasonably clever, you know what *should* happen *if all goes well*: the second bucket should end up with a volume of water identical to the volume of water originally present in the first bucket. So, based on the volume of water in the first bucket, you know how much water the second bucket will have *if all goes well*. And based on the volume of water in the second bucket, you know how much water was in the first bucket *if all went well*. Why is this important? Because you can now determine whether there's a leak or not and quantify how much water you lost by computing the ratio between the two volumes of water. You can do this because you know the theoretical relation between the two volumes *under ideal conditions*—you know what should happen if all goes well.

Metaphorically, you can think of *meta-d'* as an indicator estimating how much sensory evidence should be in the bucket of perception based on the sensory evidence in the bucket of metacognition. By comparing the subject's actual *d'* to *meta-d'*, one can determine whether the subject made optimal metacognitive use of the evidence available for the Type-1 decision or not. If *meta-d'* = *d'*, the subject *is* a metacognitively ideal observer. All the sensory evidence available for her Type-1 decisions was also available for her metacognitive decisions. On the other hand, if *meta-d'* < *d'*, the subject is metacognitively suboptimal: some sensory evidence available for the Type 1 decisions was not available for the Type 2 decisions (or was not appropriately used for metacognitive computations)[10]. Some evidence in the bucket of perception didn't make it to the bucket of metacognition.

Since the value of *meta-d'* is in the same units as *d'*, one can obtain the ratio *meta-d'/d'*, which indicates the participant's metacognitive efficiency (Maniscalco & Lau, 2012). This is the M-ratio. An M-ratio of 1 indicates that the subject systematically uses all the sensory information available in the Type 1 task to perform the Type 2 task; namely, the subject performs the Type 2 task like an ideal metacognitive observer. A lower M-ratio indicates metacognitive suboptimality—the lower the ratio the more suboptimal. The M-ratio thus provides a way of evaluating a subject's metacognitive performance, relative to her performance on the Type 1 task—her metacognitive efficiency, based on how she behaves compared to an ideal metacognitive observer. That's all for our non-technical introduction to metacognitive indicators. I'll add more details along the way. For now, let's come back to consciousness.

## 3. Interpreting metacognitive indicators for consciousness research

### 3.1. The basic idea

[10] In some circumstances, meta-d' > d' (see e.g. Charles et al. 2013). There are several possible explanations for this. Perhaps sensory evidence continues to accumulate after the Type-1 decision, meaning that more evidence is available for the Type-2 decision than for the Type-1 decision (Pleskac & Busemeyer, 2010; Moran et al. 2015; Murphy et al. 2015). Another explanation is that the metacognitive decision is influenced by non-sensory factors, such as the motor fluency of the report of the Type-1 decision (Fleming et al. 2015; Fleming & Daw, 2017; Gajdos et al. 2019). Finally, Miyoshi & Lau (2020) have shown that, when the SDT assumption of equal variance is rejected, meta-d' > d' is expected for a metacognitive observer who ignores decision-incongruent information.

Let me describe the perfect world for using confidence-based procedures in consciousness research. In this world, subjects are sensitive to the correctness of their responses *only when* the sensory activity driving these responses is *conscious* sensory activity. Their metacognitive system is completely blind to *unconscious* sensory activity, but whenever some *conscious* sensory activity occurs, it is automatically available to the subjects' metacognitive systems. This happens through some magic—or as philosophers call it, 'acquaintance'—such that the metacognitive system performs its computations without inefficiencies based on uncorrupted inputs. Since, in this world, conscious sensory activity is always available for the subjects to *think* that they're currently perceiving such and such features, the contents of consciousness never 'overflow' from metacognition (Block, 2007, 2011). Finally, subjects never experience visual hallucinations in this world—I'll explain why that's important later (see Objection 6 as well).

We just imagined a world where unconscious sensory activity is metacognition's only blind spot—its Achilles' heel. An observer in this world is metacognitively ideal if, and only if, she *consciously* perceives all the features she perceives. Now, remember: *Meta-d'* quantifies the Type-1 sensitivity that a subject would have had, had she been a metacognitively ideal observer. This means that, in this world, *meta-d'* quantifies the sensitivity that the subject would have had, had she consciously perceived all the features she perceived. It follows that *Meta-d'–d'* is the difference between the subject's actual sensitivity, and the sensitivity she would have had, had she consciously perceived all the features she perceived. If *meta-d'* = *d'*, all the features she perceived were *consciously* perceived. And if *meta-d'* < *d'*, she perceived some features unconsciously. Indeed, *had* the subject consciously perceived all the features she perceived, *meta-d'* would have been equal to *d'*. Similarly, in this world, the M-ratio is an indicator of consciousness. An M-ratio of 0.7 means that 70% of the subject's sensitivity is explained by her consciously perceiving the relevant feature. That's a pretty nice result, even if it only obtains in an ideal world.

We all agree that our world is far from this world. But it is important to understand *how far*, and what that means for using confidence-based procedures in consciousness research. The fact that we are very far from this world can have two kinds of consequences for confidence-based procedures. It could affect their validity, their accuracy, or both.

Validity is whether a procedure actually measures the intended attribute (Borsboom et al. 2004). Meanwhile, following classical test theory, I define accuracy as the correlation between a test score and the true score (Lord & Novick, 1968). What matters for us is what the distinction between validity and accuracy conveys: there is a difference between being bad at measuring something and not measuring it at all. Using my thermometer to measure the GDP of Malawi is not valid. Using it for measuring temperature is valid. And given that using it is valid for measuring temperature, its accuracy can be assessed with respect to that property[11]. Perhaps my thermometer always overestimates the true temperature by 14°C. This doesn't mean that it doesn't measure temperature. It's just that it does a bad job at it. In the same way, there's a difference between confidence-based procedures being *invalid* for assessing consciousness; and confidence-based procedures being *inaccurate* for assessing consciousness.

In the imaginary scenario described above, I set out conditions guaranteeing the validity and accuracy of confidence-based procedures. These conditions are as follows:

> Consciousness-Selectivity: Unconscious sensory activity is a source of metacognitive inefficiency.
>
> Optimal Metacognition: Observers have optimal metacognitive sensitivity for perceptual decisions based on conscious sensory activity.
>
> No overflow: All conscious sensory activity is available for metacognitive decisions.
>
> No hallucination: Sensory activity driving false alarm responses is not conscious.

Consciousness-Selectivity guarantees the *validity* of confidence-based procedures: differences in consciousness lead to differences in metacognitive efficiency. It also tells us how that difference should manifest: metacognitive efficiency is lower when subjects unconsciously perceive visual features compared to when they consciously perceive them.

Optimal Metacognition, No Overflow, and No Hallucination guarantee the *accuracy* of confidence-based procedures. Together, these conditions guarantee that a decrease in metacognitive sensitivity can be *fully attributed* to unconscious perception. Failing to satisfy Optimal Metacognition and No Hallucination leads to variations in metacognitive sensitivity that do not track differences in perceptual consciousness—or *construct-irrelevant variance* (Markus & Borsboom, 2013, p.55). Failing to satisfy No Overflow leads to differences in consciousness that are not tracked by variations in metacognitive sensitivity—or *construct-underrepresentation* (Markus

[11] I assume with Borsboom et al. (2004) that validity is a precondition for a psychometric property such as accuracy as defined here. It makes no sense to ask how accurately my thermometer measures the GDP of Malawi because it simply doesn't measure it at all. One could perhaps answer that my thermometer *does* measure the GDP of Malawi, but with zero accuracy. This response, however, has the unfortunate consequence that everything is a measurement of everything else, albeit with zero accuracy.

& Borsboom, 2013, p.55). In these conditions, while differences in metacognitive sensitivity might track differences in perceptual consciousness (assuming Consciousness-Selectivity), confidence-based procedures are inaccurate. One will infer differences in consciousness between conditions where there are none; and one will conclude that there is no difference in consciousness between conditions that differ with respect to consciousness.

Optimal Metacognition and No Hallucination are obviously not satisfied in our world (Shekhar & Rahnev, 2020; Rahnev & Denison, 2018). The jury is still out for No Overflow. Ultimately, confidence-based procedures are probably quite inaccurate. Nevertheless, I will argue that we can identify conditions in which confidence-based procedures are accurate *enough* for our practical and scientific purposes (Michel, 2021). This discussion will have to wait though. Establishing the *validity* of confidence-based procedures is a more pressing matter. If Consciousness-Selectivity is not satisfied, confidence-based procedures are not just inaccurate, but simply hopeless. Anyone relying on confidence-based procedures has to accept Consciousness-Selectivity, or at least some version of it. So let me defend it now before coming back to the issue of accuracy.

## 3.2. Are confidence-based procedures valid?

A good explanation for the widespread use of confidence-based procedures in consciousness research is their *face validity*. The following brute intuition seems widely shared: all other things being equal, if I feel like I clearly saw a stimulus, I will be more confident in my perceptual decisions about that stimulus than if I feel like I didn't see anything at all. If this is correct, variations in consciousness lead to variations in confidence. And if that's the case, confidence-based procedures are valid.

This 'brute intuition argument' is unlikely to convince anyone who is skeptical about the validity of confidence-based procedures to begin with. So, let me hesitantly complement it with a philosophical argument before turning to empirical arguments.

One can appeal to the relation between consciousness and epistemic justification to argue for the validity of confidence-based procedures. Suppose that I consciously perceive a square, and thus respond that the stimulus is a square. My conscious perceptual state *justifies* my belief (and my response) that the stimulus is a square.

Many epistemologists argue that unconscious perceptual states do not likewise justify beliefs (Byrne, 2016; Campbell, 2002; Huemer, 2001, 2006; Johnston, 2006; Smithies, 2019), or at least not *to the same extent* as conscious perceptual states (Silins, 2011). In particular, Smithies (2019) argues that whatever provides justification must be introspectively available. If this epistemological view—or its watered-down version in terms of degrees of justification—is correct, conscious and unconscious perceptual states do not have the same justificatory strengths. Assuming that confidence tracks justification, one should thus accept that confidence-based procedures are valid. Confidence follows justification; justification follows consciousness; and thus, differences in consciousness should result in differences in confidence.

The role of consciousness for justification is nevertheless a matter of debate (Berger, 2014, 2020; Jenkin, 2020; Siegel, 2017). Without entering too deep into this issue, take the following case from Berger et al. (2018):

> Consider the ordinary experience of sitting in a crowded place … and suddenly feeling as though you are being watched. You might quickly look up to lock eyes with someone nearby. A reasonable explanation is that you unconsciously saw the person looking at you; this unconscious perception then caused, and likewise justified, the conscious belief that you were being watched (Berger et al. 2018; p.571)

At the cost of doing a disservice to my cause, I tend to side with Berger et al. here. The argument above is far from decisive—even though one could still maintain an asymmetry in *justificatory strength* for conscious and unconscious perceptual states, in which case the argument still goes through.

While it is important to note that one's view on this epistemological issue has consequences for the validity of confidence-based procedures, I now leave this debate aside in order to focus on empirical evidence supporting the Consciousness-Selectivity condition.

A central piece of evidence comes from a study by Vlassova et al. (2014). Participants discriminated the direction in which a fraction of randomly moving dots moved coherently, and indicated their confidence (Figure 6A). Before target presentation, subjects saw salient moving dots in one eye, and either randomly moving or dots coherently moving to the left or to the right in the other eye. This masking—which is akin to continuous flash suppression—was effective: subjects were at chance discriminating whether the masked dots moved coherently or not. Yet, this episode of unconscious sensory registration had a significant impact on the discrimination task: it improved sensitivity to motion direction when masked and unmasked dots had congruent motion (Figure 6B). The key result is that despite their performance benefiting from this 'bonus' of unconscious sensory activity, the participants' confidence ratings didn't budge (Figure 6C), with identical *meta-d'* values between the two conditions (Figure 6D). Performance benefited from unconscious sensory activity. *Meta-d'* was unaffected by it. That's a point for Consciousness-Selectivity.

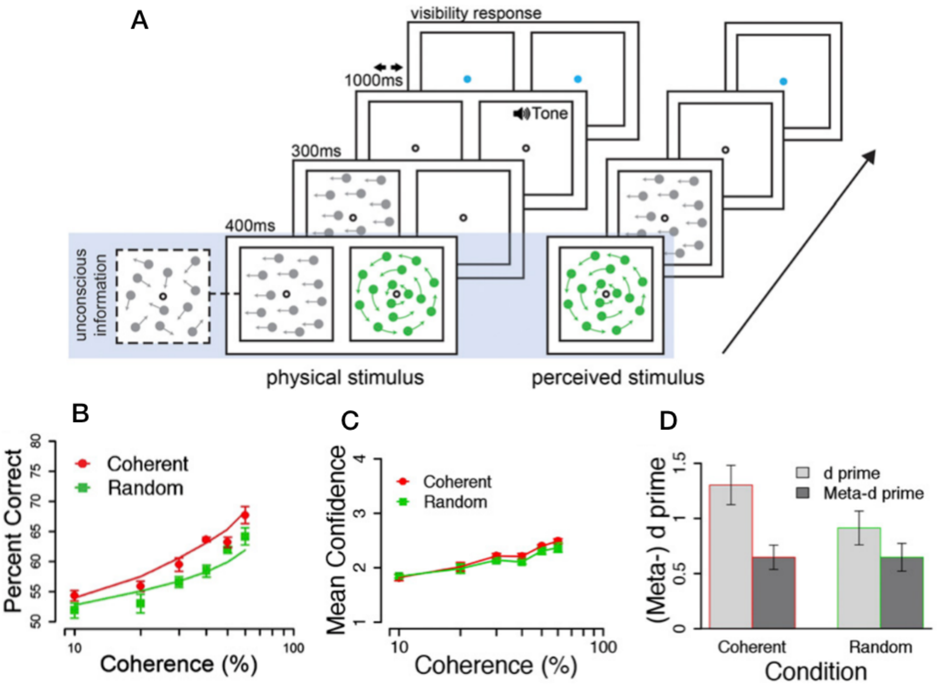


**Figure 6.** Adapted from Vlassova et al. (2014). A. Experimental paradigm. See main text. B. Type-1 performance following either randomly moving or coherently moving masked dots. C. Mean confidence in the two conditions. D. *d'* and *meta-d'* in the two conditions.

Some more evidence comes from blindsight subjects' performance in post-decision wagering tasks (Persaud et al. 2007; Persaud et al. 2011). Blindsight is characterized by residual visual abilities in the absence of reported visual awareness following lesions to the primary visual cortex (Weiskrantz, 2009)[12]. In post-decision wagering, a bet on the accuracy of one's decisions replaces confidence ratings (Persaud et al. 2007). Two studies with blindsight patient GY are relevant here. Persaud et al. (2007) showed that when GY is aware of stimuli, his bets tend to track the accuracy of his decisions. This is true both when stimuli are presented in his sighted field, but also for stimuli he reported being aware of in his 'blind' field. Meanwhile, the rate of

[12] Absence of visual awareness in blindsight can be established independently of confidence-based procedures (Azzopardi & Cowey, 1997; Persaud & Cowey, 2008; Michel & Lau, 2021), thus allowing an independent evaluation of those procedures in blindsight patients (Michel, 2021).

advantageous bets falls dramatically for subthreshold stimuli, despite performing well above chance on a discrimination task (Figure 7A). Persaud et al. (2011) replicated this result, but carefully matched performance between the sighted and 'blind' fields, thus ensuring that the difference in metacognitive sensitivity could not be attributed to a difference in Type-1 performance (Figure 7B). Importantly, these results cannot be explained by loss-aversion, or a conservative criterion for betting high in the blind field: in both studies, GY was willing to bet on his blind field performance (Figure 7C). The difference between sighted and 'blind' hemifields is a difference in metacognitive *sensitivity*; not metacognitive bias (of note: adopting the Naïve view would lead one to the opposite conclusion). These results confirm the prediction that differences in consciousness lead to differences in metacognitive sensitivity.

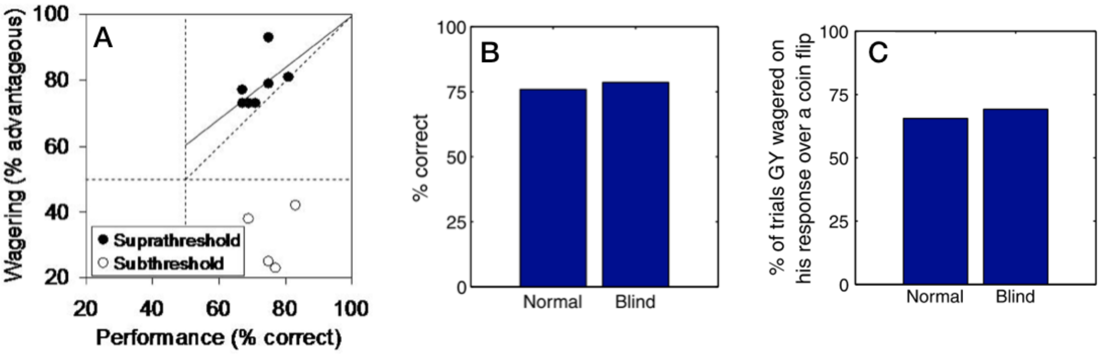


**Figure 7.** A. Wagering compared to performance in aware and unaware conditions. Each point represents data from one block of trials. Performance is similar between the two conditions, but GY does not wager advantageously for subthreshold stimuli (Source: Persaud et al. 2007; Supplementary Figure). B. Matched performance between the 'blind' and sighted visual fields. C. GY is willing to bet on his performance in the 'blind' visual field (Source: Persaud et al. 2011).

Another prediction following Consciousness-Selectivity is that participants should exhibit poor metacognitive sensitivity to the difference between misses and correct rejections in detection tasks. That's because, presumably, participants don't experience anything in both cases. So, if metacognition tracks conscious sensory activity, one should expect an asymmetry between metacognitive sensitivity for judgments of absence, and metacognitive sensitivity for judgments of presence. This is precisely what we observe: metacognitive sensitivity is notoriously bad for judgments of absence in detection tasks (Kellij et al. 2021; Mazor et al. 2021; Mazor & Fleming, 2020; Meuwese et al. 2014)[13].

[13] The exception is when participants miss stimuli because of manipulations of attention (Kanai et al. 2010). In that case, participants can presumably monitor their own attentional states to distinguish between misses and correct rejections, thus increasing metacognitive sensitivity. This is not a problem for

Third prediction: if confidence-based procedures are valid, their outcomes should generally correlate with those of other valid procedures, such as visibility-based procedures. While some dissociations are observed (see Section 4), outcomes of confidence-based and visibility-based procedures generally correlate to a very large extent (Peters & Lau, 2015; Rausch et al. 2021; Sandberg et al. 2010; Zehetleitner & Rausch, 2013) (See Objection 4 for more radical dissociations). After comparing confidence-based and visibility-based procedures, Zehetleitner and Rausch (2013) concluded: "there was a considerable association between the two ratings that were required after each trial, indicating that the patterns of the ratings are quite similar" (p.1423). In Section 4, I will argue that dissociations observed at the margins between outcomes of visibility-based and confidence-based procedures can be explained away in terms of differences in accuracy. One can thus maintain that the large shared variance in outcomes is explainable in terms of shared validity, while the small unshared variance in outcomes is due to differences in accuracy.

Here is a final, related argument. The obvious validity of visibility-based procedures probably stems from the fact that whatever cognitive system produces visibility ratings only takes *conscious* sensory activity as input, or gives different outputs depending on whether it takes *conscious* or *unconscious* sensory activity as input. Something like the Consciousness-Selectivity condition is likely correct for visibility ratings. Now, there is considerable evidence indicating that visibility ratings and confidence ratings are outputs of largely overlapping (meta)cognitive systems in prefrontal cortex (Dehaene et al. 2001; Del Cul et al. 2009; Fleming et al. 2014; Mashour et al. 2020; Mazor et al. 2021; Michel & Morales, 2019; Rahnev et al., 2016; Rounis et al. 2010; Shekhar & Rahnev, 2018; van Vugt et al. 2018)[14]. In the same way, similar factors affect visibility and confidence. For instance, pre-stimulus neuronal excitability leads to similar biases in confidence and visibility (Benwell et al. 2017; Samaha et al. 2017). Ultimately, the systems generating confidence and visibility judgments could simply be one and the same metacognitive system performing different computations over the same inputs (Fleming, 2019; Rausch et al. 2021). Since whatever system generates visibility reports has to be consciousness-selective, this

the Consciousness-Selectivity condition since the increase in metacognitive sensitivity in this case presumably does not come from metacognition tracking unconscious sensory activity.

[14] I am not assuming that the prefrontal cortex is relevant for consciousness (See Malach (2022) and Michel (2022b) for reviews on this issue). Instead, I argue that systems responsible for *evaluating* conscious visibility are dependent on prefrontal cortex activity, and these systems overlap with the metacognitive systems responsible for confidence judgments.

large mechanistic overlap should bring inductive support for the idea that the metacognitive system that outputs confidence ratings is consciousness-selective as well.

Future work should aim to further validate confidence-based procedures. The issue is far from settled. But I suggest that the current evidence—adding to the initial face validity of these procedures—indicates that confidence-based procedures are indeed valid for assessing consciousness. Whether a mental state is conscious or not *does* make a difference for metacognition. This does not necessarily mean that confidence-based procedures are accurate. And this does not mean that they are more accurate than alternative procedures either. I discuss these issues in the next sections.

### 3.3. Are confidence-based procedures accurate?

Even if unconscious perception is indeed a source of metacognitive inefficiency, it is only one of many. Here, I adopt Shekhar & Rahnev's (2020) classification:

> existing sources of metacognitive inefficiency could be understood based on two key considerations: (i) does the corruption arise from systematic (predictable) or nonsystematic (random) causes; and (ii) is the corruption due to the input to the confidence computation or due to the confidence computation itself. Taken together, these two dimensions create four categories of metacognitive inefficiency. (p.5)

As I characterized it, unconscious perception results in a *systematic input failure*—"the system responsible for generating confidence ratings does not have access to the same type or quality of information as the system making the primary decision" (Shekhar & Rahnev, 2020, p.7). But one could also categorize it as a *systematic computation failure*. Perhaps metacognitive systems *do* take the relevant unconscious sensory activity as input, but exclude it for some reason.

There are many other systematic sources of metacognitive inefficiency (for a review, see Shekhar & Rahnev, 2020). For instance, metacognition exhibits a positive evidence bias (Koizumi et al. 2015; Peters et al. 2017; Michel & Peters, 2020). Decision congruent and incongruent evidence are not given equal weight, thus leading to suboptimal metacognitive decisions.

As for non-systematic sources of metacognitive inefficiency, scientists generally categorize them as 'metacognitive noise': "random, nonsystematic noise in the confidence ratings that is not present in the perceptual decision" (Shekhar & Rahnev, 2020, p.8)[15].

[15] An experiment by Bang et al. (2019) provides strong evidence for the existence of metacognitive noise. A counterintuitive prediction if metacognitive noise exists is that decreasing *sensory* noise should *decrease*

These sources of metacognitive inefficiency are confounding factors, or sources of detection error. Unconscious perception is not the only source of metacognitive inefficiency. Any decrease in metacognitive efficiency observed between two conditions cannot be directly attributed to unconscious perception. For this reason, experimenters have to combine confidence-based procedures with experimental procedures allowing them to secure the inference from metacognitive inefficiency to unconscious perception (Staley, 2004, 2020). I identify three main ways of doing so: *ceteris absentibus*, *ceteris paribus* and *ceteris neglectis* conditions[16].

One way of controlling for confounding factors is to eliminate them. Unconscious perception is left as the only source of metacognitive inefficiency, *ceteris absentibus*—all other factors being absent. Of course, completely *ceteris absentibus* conditions are hard to get. But we can strive for conditions that are as *ceteris absentibus* as possible. For instance, memory failures could constitute a systematic source of metacognitive inefficiency (See Objection 5). Experimenters can get closer to *ceteris absentibus* conditions with respect to this confounding factor by using a scale allowing participants to report the Type-1 and Type-2 decisions at the same time.

Confounding factors that can't be eliminated should, as much as possible, be made equal between different conditions of an experiment, thus creating *ceteris paribus* conditions—all other things being equal. Achieving this requires creating conditions that only differ in conscious visibility, and comparing metacognitive efficiency between those two conditions. Given *ceteris paribus* conditions, one can assume that confounding sources of metacognitive inefficiency will be similar enough between the relevant conditions, leaving the difference in consciousness as the main factor explaining the difference in metacognitive efficiency.

Experimental conditions are never completely *ceteris absentibus* and *ceteris paribus*. But striving to create the right experimental conditions might lead to *ceteris neglectis* conditions—where all other things are negligible. The accuracy of confidence-based procedures ultimately depends

metacognitive efficiency. Essentially, this is because the effect of decreasing sensory noise is relatively more important for the Type-1 than Type-2 decision. This is precisely what Bang et al. (2019) observed, thus validating models that posit metacognitive noise.

[16] My analysis follows Boumans and Morgan's (2001) analysis of *ceteris paribus* conditions in economics (see also Boumans, 1999; Morgan, 2013)

on how negligible these other factors really are[17], and how detrimental their presence is to the goal of the experiment.

Because accuracy depends on specific experimental conditions, one can only evaluate it on a case-by-case basis by determining whether unconscious perception can be legitimately claimed as a source of metacognitive inefficiency. Either because, *ceteris absentibus*, all confounding factors have been eliminated. Or *ceteris paribus*, all confounding factors are matched between the purportedly conscious and purportedly non-conscious conditions. Or *ceteris neglectis*, all remaining confounding factors should only have a negligible influence that cannot account for the relevant effect size.

Establishing these conditions is extremely difficult. But asking for perfect accuracy is an unreasonable standard. All measurement is subject to measurement error. The relevant question is whether our procedures are accurate *enough* given our practical and epistemic goals. What is enough? A procedure is accurate enough for a given goal if a more accurate procedure would have provided the same outcome. That is, a procedure is accurate *enough* when its inaccuracy does not undermine its epistemic and practical functions (Michel, 2021; see also Elgin, 2017).

Let me illustrate. Galileo's telescope-based procedure for detecting craters on the moon was extremely inaccurate compared to the procedures we would use today to achieve the same goal. But it was accurate *enough*. Galileo would have reached the same conclusion using a more accurate procedure. The inaccuracy of his procedure did not undermine his goals[18]. Similarly, confidence-based procedures are most certainly inaccurate. But one can create conditions where they are accurate *enough* for our scientific goals. Again, whether this is the case or not for confidence-based procedures can only be decided on a case-by-case basis.

[17] In some cases, one can model the influence of the remaining confounding factors on the procedure in order to determine whether or not they are sufficient to account for the observed difference in metacognitive inefficiency between two conditions (Michel, 2021). This assumes that one can model the relevant confounding factors, but models including Type-2 noise have already been used for the purpose of identifying unconscious perception (Peters & Lau, 2015), and the factors influencing metacognitive efficiency are increasingly well understood (Shekhar & Rahnev, 2020).

[18] Measurement and detection outcomes are scientific representations (van Fraassen, 2008). Just as in the case of pictorial representations, accuracy is partly goal-dependent. Borrowing an example from van Fraassen (2008), a cartoon picturing Margaret Thatcher as a dragon is inaccurate if one uses it to know what she looked like; but it might be an accurate depiction of her personality. Galileo's detection outcome was accurate enough for his purposes, but would not have been accurate enough for other purposes.

In the next section, I argue that confidence-based procedures are likely *more* accurate than alternative procedures in many cases. Then, in Section 5., I use the analysis developed so far to answer some common objections against confidence-based procedures.

## 4. Confidence-based procedures are better than other procedures

The main alternative to confidence-based procedures are visibility-based procedures, either with simple visibility ratings such as 'seen'/'not seen', or with PAS ratings, which typically include the following responses: 'No experience', 'brief glimpse', 'almost clear image' and 'absolutely clear image' (Ramsoy & Overgaard, 2004; Sergent & Dehaene, 2004). I will not discuss the main problems with those procedures here (Michel, 2019, 2022a), but simply highlight where I believe that confidence-based procedures fare better.

Confidence ratings reflect the subjective probability of being correct. For this reason, the outcomes of confidence-based procedures can be meaningfully compared across tasks. This is the case across modalities (Deroy et al. 2016; de Gardelle et al. 2016; Faivre et al. 2017), and domains such as perception and memory (McCurdy et al. 2013; Morales et al. 2018). Confidence-based procedures can also be used across species, allowing consciousness researchers to obtain metacognitive indicators of consciousness in non-human animals (Kepecs & Mainen, 2012; Kiani & Shadlen, 2009; Smith et al. 2014). Other procedures are tied to specific modalities, their outcomes cannot be compared across tasks, and they cannot be used in non-human animals. So, confidence-based procedures are better than alternative procedures.

Another advantage of confidence-based procedures is that they probably assess consciousness of the *task-relevant features* of the stimuli, namely, those features on which Type-1 performance depends (Michel, 2022a). Why? Because consciously perceiving features that are completely irrelevant for performing the task shouldn't make you more confident that your decisions are correct. Meanwhile, we have reasons to believe that visibility-based procedures do not specifically track consciousness of the task-relevant features, but consciousness of *something*—whatever that something happens to be. Let me illustrate with a scenario inspired from an experiment by Koivisto & Neuvonen (2020).

Suppose that your task is to discriminate the orientation of masked Gabor patches with random colors. Color is completely task-irrelevant. Suppose also that, on some trials, you definitely see *something*, you see the *color* of the stimulus, and yet, do not consciously perceive the *orientation*. Confidence-based procedures and visibility-based procedures are likely to lead to different verdicts here.

Using a confidence rating scale, you might answer 'guess' as long as you didn't consciously perceive the *orientation*. After all, color is task-irrelevant. Seeing it shouldn't make you more confident in your judgment about orientation.

Things are different with a visibility or PAS scale. You did see the stimulus. So, you might answer 'seen', 'brief glimpse' or 'almost clear image', even if you didn't consciously perceive the task-relevant feature (orientation) (Dienes & Seth, 2010). These Type-2 responses might in turn lead experimenters to conclude that you were conscious of the *orientation*, when all you consciously saw were task-irrelevant features. Visibility-based procedures might lead researchers to commit the *criterion content fallacy*: the fallacy of concluding that you consciously perceived *task-relevant features* on the basis of the mere evidence that you saw *something* (Michel, 2022a).

Cases like this one are probably quite common. Complete suppression with visual suppression techniques is the exception rather than the rule (Breitmeyer, 2014). In most cases, some stimulus features remain to be consciously seen even when the *task-relevant* features are fully suppressed. For instance, using visual masking, Koivisto & Neuvonen (2020) showed that participants can report being conscious of a task-irrelevant feature (e.g. color) without being conscious of the task-relevant feature (e.g. orientation). Stober et al. (1978) and Breitmeyer et al. (2006) showed that metacontrast masks can sometimes suppress stimulus brightness, but not contour features, and vice versa. Kim & Shong (2021) also report independent access to various stimulus features in masking—superordinate vs. basic categories; local vs. global features; and low vs. high spatial frequency. Similarly, with binocular flash suppression techniques, Gelbard-Sagiv et al. (2016) showed that participants can be aware of the color or location of a suppressed face without being able to discriminate its identity. Hong & Blake (2009) obtained the same result with color and orientation (See also Pournaghdali & Schwartz, 2020).

Asking subjects to provide visibility ratings essentially requires them to collapse a multi-dimensional percept of task-relevant and task-irrelevant features into a single 'visibility' dimension (Koster et al. 2020; Sackur, 2013). Without clear instructions, the participants are free to determine how to weigh these different dimensions. Confidence ratings, on the other hand, are likely to track consciousness of the *task-relevant features*, since perception of those features is what matters for successful Type-1 performance[19].

This difference between confidence-based and visibility-based procedures has been confirmed across several studies by Rausch et al. (Rausch et al., 2015, 2018; Rausch & Zehetleitner, 2016; Zehetleitner & Rausch, 2013). Confidence ratings tend to track the accuracy of the decision. Visibility ratings tend to track the objective strength of the stimulus (e.g. contrast), even when the participants' responses are incorrect (Rausch & Zehetleitner 2016). Rausch et al. (2021) write: "A comparison between subjective visibility and decisional confidence revealed that visibility relied more on the strength of sensory evidence about features of the stimulus irrelevant to the identification judgment and less on evidence for the identification judgment." (p.3311). Additional evidence for this claim comes from the fact that participants often report 'brief glimpse' when using the PAS even when Type-1 performance does not differ from chance, thus indicating that PAS reports dissociate from the conscious perception of task-relevant features (Mazzi et al. 2016; Jimenez et al. 2019).

So, visibility ratings and the PAS track consciousness of *something*. Confidence-based procedures track consciousness of the *task-relevant features* of the stimulus. This could explain dissociations between PAS ratings and confidence-based procedures reported by Sandberg et al. (2010) and Wierzchoń et al. (2014). These studies showed that the PAS is more 'exhaustive' than confidence-based procedures—PAS ratings showed better correlations with Type-1 performance than confidence ratings. PAS was also more 'sensitive'—performance was lower when participants reported with the lowest PAS category compared to when they reported with the lowest confidence category. For instance, Sandberg et al. (2010) report that accuracy when participants used the lowest report-category was 27.9% for the PAS (just above chance level in this task), and 36.6% for confidence ratings. Does that mean that confidence-based procedures overestimate unconscious perception? Not necessarily.

The difference can be explained by the fact that PAS-based and confidence-based procedures do not assess exactly the same thing (Michel, 2022a). As above, suppose that you

[19] This is not to say that confidence-based procedures are not influenced by conscious perception of task-irrelevant features at all. Sensory evidence strength for task-irrelevant features could be used as a proxy for the strength of sensory evidence for task-relevant features, which, in turn, provides a good indication of Type-1 performance. What matters is that confidence-based procedures are *relatively* less influenced by conscious perception of task-irrelevant features than alternative procedures (Rausch et al. 2021).

perform a discrimination task with masked geometrical shapes. You sometimes get the feeling that *something was there*, but you subjectively feel like you didn't see the *shape*. In those conditions, you might report seeing a 'brief glimpse' with the PAS—you definitely saw *something*. And you might answer 'guess' with the confidence scale—after all, you didn't see the *shape*. Assuming your response was correct, the PAS-based procedure would categorize this trial as a Type-II hit, but the confidence-based procedure would categorize it as a Type-II miss[20]. This could lead to the dissociations observed by Sandberg et al. (2010) and Wierzchon et al. (2014).

In sum, what consciousness researchers want to know is whether participants are conscious of the visual features used to perform the Type-1 task. Not something else. But precisely *that*. Confidence-based procedures deliver. Visibility-based procedures don't.

The final advantage of confidence-based procedures is that we increasingly understand how confidence ratings are generated. Ideal metacognitive observer models have paved the way for the staggering progress seen in the study of metacognition in the past fifteen years by providing a clear baseline against which to judge the participants' behaviors (Fleming, 2020). This improved understanding of metacognition provides us with the opportunity to understand what *could* constitute a confounding factor on a case-by-case basis. While it is not always guaranteed that these factors can all be mapped, at least we have an idea of what could go wrong. Meanwhile, the study of the way in which visibility judgments are generated is lagging behind the study of confidence judgments. We do not currently have a good idea of the factors that could constitute confounding factors for visibility judgments, since we do not currently understand how those ratings are generated. It is thus difficult to evaluate the extent to which visibility-based procedures might be affected by systematic sources of error (although see Anzulewicz et al. 2020; Siedlecka et al. 2019, 2020; Skora et al. 2021).

In sum, confidence-based procedures can be generalized across modalities, tasks, and species. Visibility-based procedures can't. Confidence-based procedures are more likely than visibility-based procedures to assess consciousness of the *task-relevant features*, rather than

[20] In a study by Szczepanowski et al. (2013) in which subjects had to discriminate between fearful and neutral faces, metacognitive sensitivity, measured with AUROC2, was slightly higher when subjects used confidence ratings compared to PAS ratings. As noted by Zehetleitner and Rausch (2013), both Sandberg et al. (2010) and Wierzchoń et al. (2014) estimated metacognitive sensitivity with logistic regression, which could also partly explain the discrepancy between these results, as well as those obtained in a series of studies by Rausch & Zehetleitner (see below; and see Rausch and Zehetleitner (2017) on the problems with logistic regression as a way to estimate metacognitive sensitivity).

consciousness of *something*. This also explains previously reported dissociations between PAS and confidence-based procedures. Finally, we currently don't know which factors could influence visibility-based procedures, making it difficult to control for confounding factors. Meanwhile, the way in which confidence ratings are generated is increasingly well understood, allowing us to identify and control for confounding factors.

## 5. Eleven objections and responses

In this section I do my best to address eleven objections to the use of confidence-based procedures. Some of these objections rely on incorrect assumptions about the ways in which confidence-based procedures are used—or in any case, how they *should* be used if one is to turn confidence ratings into valid indicators of consciousness. Other objections create genuine problems for confidence-based procedures. Hopefully, confidence-based procedures can still be valid and accurate enough despite these problems.

*Objection 1: There's a double dissociation between consciousness and confidence ratings*

Subjects can feel confident without being conscious of targets, and conscious of targets without being confident. Because of this double dissociation, confidence ratings do not indicate anything about consciousness. So, confidence-based procedures are doomed (Abid, 2019).

All kinds of dissociations are conceivable. Some are even possible. But they're largely irrelevant. First, confidence ratings are not interpreted as direct indicators of consciousness. What matters is not absolute reported confidence, but metacognitive sensitivity and efficiency.

Second, there's no reason to think that strong dissociations occur most of the time (Knotts et al. 2020). And, again, all we care about is being accurate *enough*. Here's an analogy. Sensations of hot and cold doubly dissociate from temperature. But you shouldn't conclude from this that sensations of hot and cold are not good indicators of temperature. For most purposes, my sensation-based procedure for evaluating temperature is accurate enough. The same thing could be true for confidence. In and of themselves, double dissociations are not too worrying.

Finally, if the mere possibility of double dissociations were enough to discard indicators of consciousness, there would be no indicator of consciousness. One can imagine all kinds of scenarios that would lead to double dissociations with all possible indicators. One should require something more: reasons for believing that the dissociations are bad enough and common enough to prevent us from reliably interpreting indicators as indicating consciousness (Knotts et al. 2020; Michel, 2021). So far, there's no reason to believe that's the case. And there are reasons to believe that's *not* the case. As noted above, confidence and visibility ratings can fail to correlate, but when that happens, it's only at the margins (Peters & Lau, 2015; Rausch et al. 2021; Sandberg et al. 2010; Zehetleitner & Rausch, 2013). Holding that radical dissociations between confidence and consciousness reliably occur would thus lead one to reject the validity of visibility ratings. But interpreting visibility ratings as indicators of consciousness is obviously valid.

*Objection 2: There's a double dissociation between metacognitive efficiency and consciousness*

Not only do confidence ratings doubly dissociate from consciousness, but metacognitive efficiency does too. Factors that have nothing to do with consciousness can decrease metacognitive efficiency. So, the argument goes, metacognitive efficiency cannot be interpreted as indicating anything about consciousness.

The fact that in most contexts variations in metacognitive efficiency do not track differences in consciousness does not imply that one cannot interpret variations in metacognitive efficiency as indications of consciousness in other contexts. It is not uncommon for measurement and detection procedures to require carefully controlled settings for their indications to reliably track what we want them to track. Take seismometers, for instance. A seismometer in my office in New York would not reliably track earthquakes. There are just too many confounding factors here. Seismometers require carefully controlled conditions for their indications to reliably track the presence of earthquakes. In the same way, experimenters using confidence-based procedures to study consciousness have to create the right conditions for metacognitive efficiency to be informative with respect to consciousness (Section 3.3). These conditions are designed to prevent variations in metacognitive efficiency that have nothing to do with consciousness from being confounding factors.

*Objection 3: Confidence-based procedures assume the higher-order theory of consciousness*

Higher-order theories of consciousness all have in common the hypothesis that a conscious state is a mental state whose subject is aware of being in it. I don't consciously perceive an apple if I'm utterly unaware of perceiving it. A natural way of cashing out the 'aware of' relation above is in terms of mental representation. It follows that a conscious state is a mental state that is itself represented by another of the subject's mental states (Lycan, 2001). That's the main motivation for higher-order views. On the face of it, this has nothing to do with confidence. And the validity of confidence-based procedures can be assessed without saying a word about higher-order theories. In fact, Rosenthal (2019), one of the main proponents of higher-order theory, explicitly rejects confidence-based procedures. If anything, visibility-based procedures seem more in line with higher-order theories. Presumably, a visibility rating such as 'I saw the stimulus' expresses a higher-order thought. So, confidence-based procedures do not assume higher-order theories[21].

*Objection 4: Outcomes from visibility-based and confidence-based procedures sometimes dissociate. This should be interpreted as indicating that confidence-based procedures are invalid, since visibility-based procedures are valid.*

Above, I said that outcomes of visibility-based and confidence-based procedures only differ at the margins. I also explained why these dissociations might occur (Section 4). But several studies have reported strong dissociations between outcomes of visibility-based and confidence-based procedures. Two cases are particularly pressing. Charles et al. (2013) showed that metacognitive sensitivity is well above chance for stimuli that participants report not seeing. And Jachs et al. (2015) reported that metacognitive sensitivity does not significantly differ between seen, and not seen conditions. These two studies raise serious problems for confidence-based procedures.

Charles et al. (2013) presented masked digits to subjects who had to decide whether the number was higher or lower than five, either in a 'fast responding' condition (Experiment 1) or without time pressure (Experiment 2). Visibility ratings followed, and then confidence ratings ('Error' or 'Correct'). Metacognitive sensitivity was well above chance even when subjects reported not seeing the stimuli. Charles et al. also reported that *meta-d'* was higher than *d'*.

How can we account for the dissociation? Charles et al. used 'Error' as the lowest option on the confidence scale, which is quite different from 'guess'. I can, after all, be confident that my answer was wrong. That makes it plausible that participants used this response category to indicate that they changed their mind about the Type-1 response. This, combined with response

[21] Perhaps the impression that there is a link between confidence-based procedures and higher-order theories of consciousness comes down to sociology: for obvious reasons, scientists who support higher-order theories are often interested in metacognition, and have significantly contributed to the development of confidence-based procedures (e.g., Maniscalco & Lau, 2012; Fleming & Lau, 2014). It's no surprise that scientists who study metacognition for a living also use the kind of procedures they are most familiar with when it comes to studying consciousness.

bias for the 'seen' / 'unseen' judgments and the fact that Charles et al. analyzed 'seen' and 'unseen' trials separately, might account for the result.

Suppose you're told to decide as fast as possible whether a masked number is higher or lower than five. You give your response, and then select 'Unseen' because you're not sure what you saw (perhaps relative to other trials). By the time you're asked about your confidence, you've had more time to think. You answer 'Error' if you changed your mind; and 'Correct' otherwise. Since sensory evidence continued to accumulate after the Type-1 decision, more information was available for the Type-2 decision than for the Type-1 decision, thereby accounting for *meta-d'* > *d'* (Pleskac & Busemeyer, 2010; Moran et al. 2015; Moreira et al. 2018; Murphy et al. 2015). This explanation is in line with the fact that the effect largely decreased in Experiment 2, without time pressure for the Type-1 decision. In addition, given a relatively conservative bias for 'seen' judgments, 'unseen' trials analyzed in isolation might include trials in which participants were conscious of the stimuli, which could then account for above-chance metacognitive performance on (purportedly) 'unseen' trials (Schmidt, 2015; Shanks, 2017).

Anyone committed to the validity of confidence-based procedures should predict that the observed dissociation between metacognitive sensitivity and visibility would be greatly reduced if participants were required to report confidence at the same time as the Type-1 decision. Doing so would, however, require changing the confidence scale used by Charles et al. A report of the kind 'Higher than 5; Error' doesn't make any sense, but one can meaningfully report 'Higher than 5; Guess'.

Jachs et al. (2015) compared metacognitive sensitivity (with AUROC2 as well as *meta-d'*) between conditions in which participants rated stimuli as either 'seen' or 'unseen', or with PAS ratings 'Unaware' or 'Brief glimpse'. They report no significant effect of awareness on *meta-d'* across two out of three experiments, with the only significant effect being rather small. This was the case even if Type-1 sensitivity significantly differed between aware and unaware trials. Taken at face value, this result indicates a strong dissociation between awareness and metacognitive sensitivity, but also between Type-1 sensitivity and metacognitive sensitivity—which is suspicious.

This result could be explained by a combination of conservative criterion for the awareness measure and task instructions leading to an artificial dissociation between metacognitive sensitivity and awareness. Jachs et al. write: "Participants were instructed that confidence ratings should be conceived in a relative manner; accordingly, observers were instructed that the full range of these relative confidence estimates ought to be equally used both on aware and unaware rated trials." (p. 270-271). This instruction is a recipe for creating a dissociation between metacognitive sensitivity and awareness.

Suppose you're given this instruction. You consciously perceive the stimulus, and you're quite sure you saw the orientation, but you're *relatively less confident* compared to other trials where you reported being 'aware' of the stimulus. In that case you might answer 'guess', in line with the instructions, thus increasing your Type-2 miss rate. Similarly, suppose you report 'unaware'. You're not sure that you saw anything. But you're still *relatively more confident* compared to other trials where you reported 'unaware'. In that case you might answer 'sure', which could lead to an increased Type-2 hit rate (if your Type-1 response was indeed correct). These instructions could thus artificially decrease metacognitive sensitivity on 'aware' trials while increasing it on 'unaware' trials. In turn, this could explain the lack of effect of awareness on metacognitive sensitivity. Proponents of confidence-based procedures should predict that awareness will have an effect on metacognitive sensitivity if the subjects are instructed to answer 'guess' when they think their response was as good as flipping a coin, irrespective of their reported awareness[22].

*Objection 5: Memory failures cause metacognitive inefficiency, not unconsciousness.*

Another worry is that memory failures could be a pervasive source of systematic metacognitive inefficiency. As the argument goes, metacognitive inefficiency does not indicate unconsciousness, but simply memory failure. This objection applies to all procedures. This includes 'objective' procedures that directly rely on *d'* as an indicator of conscious perception. It's not a problem for confidence-based procedures in particular. In addition, it is difficult to see how this objection fits with the finding that sensory evidence continues to accumulate after the Type-1 decision, thus *increasing* metacognitive efficiency (Charles et al. 2013; Pleskac & Busemeyer, 2010; Moran et al. 2015; Murphy et al. 2015). Finally, as we saw above, one way of reducing this confounding factor is to ask participants to rate confidence at the same time as the Type-1 decision—which also improves reliability (Guggenmos, 2021).

[22] The instructions provided by Jachs et al. (2015) also lead to a violation of the ordinality of the confidence scale. For confidence ratings to be meaningful, 'sure' should indicate more confidence (or a higher subjective probability of answering correctly) than 'guess'. But by conditionalizing those response categories on awareness, 'guess' following an 'aware' judgment might indicate the same subjective probability of being correct as 'sure' following an 'unaware' judgment. This violation of the ordinality of the confidence scale makes it very difficult to establish a meaningful comparison of confidence judgments across 'aware' and 'unaware' trials, because the report categories simply don't mean the same thing across these two conditions.

*Objection 6: Hallucinations cause metacognitive inefficiency, not unconsciousness*

Suppose you experience vivid hallucinations of non-present targets. That might lead to incorrect Type-1 responses followed by 'high confidence' judgments, thereby increasing Type-2 false alarms and decreasing metacognitive sensitivity. The potential effect of hallucinations on metacognitive efficiency seems hard to control experimentally, thus creating a somewhat uneliminable confounding factor.

This objection is not so much of a problem if an experiment is comparing two conditions where there is no reason to believe that hallucinations should be more present in one condition compared to the other (*ceteris paribus* conditions). Still, it shows that experimenters should not blindly interpret metacognitive inefficiency as an indicator of unconscious perception. Unconscious perception should mainly result in a high Type-2 miss rate (low confidence on correct trials). Hallucinations should have the opposite effect of increasing Type-2 false alarms (high confidence on incorrect trials) (Schmack et al. 2021). So, if an experimental manipulation increases the Type-2 miss rate, one can conclude that hallucinations are probably not responsible for the associated decrease in metacognitive efficiency[23].

*Objection 7: No availability for metacognition does not mean no consciousness*

The contents of consciousness could 'overflow' the contents available for metacognition (Block, 2007, 2011). So, the argument goes, metacognitive inefficiency indicates unavailability for metacognition, not unconsciousness.

A similar argument would apply to visibility-based procedures. It could also apply in cases where one straightforwardly interprets *d'* as an indication of conscious perception. One could indeed argue that *d'* only indicates that some sensory activity is available for a response. And unavailability for a response in a discrimination or detection task does not mean no

[23] This kind of reasoning was used by Rounis et al. (2010) who observed decreased metacognitive efficiency following continuous theta burst suppression (a kind of transcranial magnetic stimulation on steroids) to the prefrontal cortex. As they note, "There are several ways in which TMS could have impaired metacognitive sensitivity. One possibility is that TMS reduced visibility for correct trials, which would amount to a kind of relative blindsight … Alternatively, TMS may have increased visibility for incorrect trials, a kind of "hallucinatory" effect." They ultimately favored the first interpretation because the main difference between pre-TMS and post-TMS conditions was an increase in the Type-2 miss rate, rather than an increase in the Type-2 false alarm rate.

consciousness. So, consciousness could overflow probed perceptual sensitivity. If 'overflowing' conscious contents are a problem, they're a problem for everyone.

But just because you can imagine there's a problem does not mean there's a problem. We should require good reasons for believing in the pervasiveness of overflowing contents before considering that their existence creates a significant problem for our procedures. After all, our *starting point* should be that most subjects, most of the time, are not metacognitively blind. Why? Because if it were really true that (negative) subjective reports can't be trusted, civilization as we know it would not exist.

Many of our daily activities require us to know when people don't experience things. The engineers who created my computer screen had to know that I don't experience any flicker with a refresh rate of 60 Hertz. My therapist has to know that I don't feel depressed. My doctor has to know that I don't feel any pain in my foot. An optometrist has to know that a patient doesn't see anything in her left visual field. My friend has to know that I don't feel hungry before ordering at a restaurant. Sports managers have to know that their players don't feel tired. Advertisers have to know that their commercials don't create a feeling of anger towards the brand they're advertising. To paraphrase Fodor (1990), if it's not literally true that engineers, therapists, doctors, optometrists, friends, managers and advertisers can know all this, it's the end of the world. The fact that all these practices are (relatively) successful and that the world has not ended (yet) indicates that all these people indeed know what they claim to know. This in turn provides a strong presumption in favor of the accuracy of the methods they use to know what they know—most of the time subjective reports. This strong presumption in favor of the accuracy of subjective reports means that those who hold that reports are systematically inaccurate should give us good arguments—reasons for accepting the pervasiveness of metacognitive blindness.

There are reasons for holding that subjects could be metacognitively blind *in some cases*. For instance, it could be that in the absence of attention negative subjective reports should not be trusted. We all know that we can fail to notice things when we are not paying attention. Overflowing contents are indeed often hypothesized to exist in precisely those conditions (Block, 2011, 2014). This debate is still very much open (Cohen et al. 2016). If you're worried about metacognitive blindness in those cases, here's a simple solution: don't use procedures relying on subjective reports *in those cases*.

Now, do we have reasons to think that metacognitive blindness is pervasive in other, ordinary contexts? No. There is no reason to think that subjects sitting in front of a screen, attending to a stimulus, and reporting immediately after it has been presented, are metacognitively blind (Bayne & Spener, 2010; Spener, 2015; Michel & Doerig, 2021). I am not denying that they *could be*. But we are entitled to a strong argument for believing that's the case. In most experiments, there is simply no reason to hypothesize pervasive metacognitive blindness. Innocent until proven guilty, your honor.

*Objection 8: Confidence-based procedures are overly inclusive because they can be based on gut feelings, or monitoring of one's own attentional states*

It could be that participants are more metacognitively sensitive than would be expected from conscious perception alone because of non-visual sources of metacognitive efficiency, such as 'gut feelings' or monitoring of one's own attentional states (Kanai et al. 2010). So, confidence-based procedures overestimate conscious perception.

An interesting case in this context is Type-II blindsight, in which patients can perform above chance on visual tasks without reporting visual experience, but nevertheless report non-visual feelings of presence and confidence (Brogaard, 2015; Macpherson, 2015; Rosenthal, 2019). These non-visual sources of metacognitive efficiency could lead us to underestimate unconscious perception when using confidence-based procedures (Michel, 2022a). While monitoring one's own attentional states certainly contributes to metacognitive sensitivity (Kanai et al. 2010), most people are not Type-II blindsight subjects, and it is currently unclear to what extent what happens in those patients also happens in healthy subjects. Because of this, it is difficult to evaluate whether most experiments can create *ceteris paribus* conditions with respect to these 'gut feelings', or how strong their influence actually is.

These sources of metacognitive efficiency *could* lead confidence-based procedures to systematically underestimate unconscious perception, but if that's the case, the effect is probably quite small. If gut feelings and attention monitoring were powerful enough and allowed participants to reliably track the accuracy of their responses independently of what they actually experienced, *meta-d'* = *d'* and *meta-d'* > *d'* would be observed much more frequently than they currently are. Instead, these patterns seem to be the exception rather than the rule, which means that in most cases these sources of metacognitive efficiency are not strong enough to overcome the effects of sources of metacognitive inefficiency. Despite this, possible instances of Type-II blindsight in healthy subjects as well as attention monitoring could constitute uneliminable sources of inaccuracy for confidence-based procedures.

*Objection 9: Confidence-based procedures are confounded by global estimates of correctness*

Morales & Lau (forthcoming) write:

> Consider a situation in which subjects have fixed their confidence in an experimental trial before even seeing the stimulus. This may happen via cognitive deduction, for example, when subjects know the base rate of the stimuli (e.g. that 70% of the stimuli are As rather than Bs). Before seeing the stimulus, because of their knowledge of the frequency of stimuli in that task, subjects may express high confidence in their answers independently from the quality of their conscious experiences. At the limit, subjects could become highly confident in their responses even if they don't see anything. For example, if they closed their eyes but knew that 70% of stimuli are of type A, when classifying the stimulus in a trial as 'A' they might express high confidence in the correctness of their response. This, of course, would clearly be a case where confidence is detached from consciousness.

This is again a case where confidence-based procedures could lead researchers to overestimate consciousness because of non-visual sources of metacognitive efficiency. Avoiding feedback during the main task of an experiment, as well as biased base rates in the presentation of stimuli might help circumvent this problem to some extent. It also seems plausible that *ceteris paribus* conditions might often be obtained with respect to this confounding factor. Other than that, there is not much we can do about this confounding factor at the moment.

*Objection 10: The M-ratio varies depending on criterion*

Shekhar & Rahnev (2021) found "clear empirical support for the notion that … *meta-d'/d'* provides a measure of "metacognitive efficiency," that is, it is independent of task difficulty" (p.58). That's good news[24]. But they also raised a significant concern: *meta-d'/d'* decreases as confidence criterion increases. As confidence ratings are more conservative, they are less accurate. So, our only measure of metacognitive efficiency is confounded by criterion location, both within subjects and across subjects (Xue et al. 2021). As Xue et al. (2021) write: "any manipulation that changes one's overall confidence level should be expected to produce a spurious change in estimated metacognitive efficiency" (p.7). This is a big problem in consciousness research, since visual suppression techniques likely lead to lower confidence levels overall in the purportedly unconscious condition. I expect that the same would be true for alternative indicators, such as visibility ratings.

[24] Although see Guggenmos (2021) for a small effect of Type-1 performance on the M-ratio.

This is another area where there isn't much we can do at the moment, until someone develops a modified version of M-ratio that could control for the effects of bias, or a new indicator that would avoid both task performance and bias confounds. Alternatively, the effect observed by Shekhar & Rahnev (2021) could be driven by incorrect assumptions about the distribution of metacognitive noise. Miyoshi et al. (manuscript) have shown that assuming logistic noise distributions could account for the effect of bias on metacognitive efficiency. More work needs to be done on this issue before we can say more.

*Objection 11: The M-ratio alone doesn't capture phenomenology: criterion shifts matter too.*

Researchers often interpret criterion effects as 'post-perceptual' in nature. But criterion shifts could sometimes reflect perceptual effects. This is the case for several perceptual illusions (e.g., Rosenthal et al. 2009; Grove et al. 2012) and perceptual effects like the stream-bounce effect (Rolfs et al. 2013), or filling-in of Gabor patches by collinear flankers (Polat & Sagi, 2007), as well as physiological effects such as the effect of pre-stimulus neuronal excitability on subsequent detection (Jin & Glickfeld, 2019; Samaha et al., 2017). It is currently unclear whether other effects on metacognition, like the 'positive evidence bias' (Michel & Peters, 2020) or 'subjective inflation' (Knotts et al. 2019), are perceptual or post-perceptual (but see Samaha & Denison, 2022; Knotts et al. 2020). SDT-inspired analyses cannot currently distinguish criterion effects that are perceptual in nature from post-perceptual effects (Witt et al. 2015). For this reason, the M-ratio, in and of itself, might fail to reflect changes in phenomenology that are instead reflected in confidence criterion changes. As a matter of fact, several studies have interpreted confidence criterion shifts as indicating that criterion effects are perceptual in nature (Gallagher et al. 2018, 2021). Future research should aim to develop analyses allowing to evaluate the respective contributions of perceptual and non-perceptual factors on confidence criterion shifts. In the meantime, experimental manipulations changing the confidence criterion without an effect on the M-ratio should not be automatically discarded as reflecting non-perceptual effects (Peters et al. 2016).

*Summary*

Confidence-based procedures are far from perfect. One should evaluate, on a case-by-case basis, whether these procedures are accurate *enough* for the task at hand. While their accuracy can certainly be questioned, the main arguments against their validity fall short of demonstrating that interpreting metacognitive indicators as indicators of consciousness is invalid. Future work should aim to further validate these procedures, as well as develop new, more accurate indicators.

## Conclusion and a path forward

Confidence-based procedures are valid for assessing consciousness. While not perfectly accurate, these procedures might often be accurate *enough*. The standard arguments against confidence-based procedures do not justify rejecting them. Despite this relatively optimistic conclusion, there is still quite a lot of work to do. Hopefully, the path forward is now clearer. One of the main goals of this research program should be to further validate confidence-based procedures. This can be done by developing experimental paradigms similar to those of Vlassova et al. (2014) and Persaud et al. (2011) to establish unconscious perception as a robust source of systematic metacognitive inefficiency. Cases in which metacognitive efficiency and awareness dissociate can be further investigated as well (Charles et al. 2013; Jachs et al. 2015). Assuming their validity, there are several outstanding issues with respect to the accuracy of confidence-based procedures. In particular, the recent discovery that metacognitive efficiency depends on metacognitive bias seems particularly important for the purposes of using metacognitive indicators in consciousness research (Shekhar & Rahnev, 2021). Further research on the nature of confidence criterion shifts is also important for future progress.

## References

Abid, G. (2019). Deflating inflation: the connection (or lack thereof) between decisional and metacognitive processes and visual phenomenology. *Neuroscience of Consciousness*, *2019*(1), 1–7.

Anzulewicz, A., Hobot, J., Siedlecka, M., & Wierzchoń, M. (2019). Bringing action into the picture. How action influences visual awareness. *Attention, Perception, and Psychophysics*, *81*(7), 2171–2176.

Azzopardi, P., & Cowey, A. (1997). Is blindsight like normal, near-threshold vision? *Proceedings of the National Academy of Sciences of the United States of America*, *94*(25), 14190–14194.

Balsdon, T., & Clifford, C. W. G. (2018). Visual processing: Conscious until proven otherwise. *Royal Society Open Science*, *5*(1), 1–16.

Bang, J. W., Shekhar, M., & Rahnev, D. (2019). Sensory Noise Increases Metacognitive Efficiency. *Journal of Experimental Psychology: General*, *148*(3), 437–452.

Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. *PLoS Computational Biology*, 5, e1000504.

Barthelmé, S., & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. P*roceedings of the National Academy of Sciences*, 107, 20834–20839.

Bayne, T., & Spener, M. (2010). Introspective Humility. *Philosophical Issues*, *20*, 1–22.

Benwell, C. S. Y., Tagliabue, C. F., Veniero, D., Cecere, R., Savazzi, S., & Thut, G. (2017). Prestimulus EEG Power Predicts Conscious Awareness But Not Objective Visual Performance. *ENeuro*, *4*(6), 1–17.

Berger, J. (2014). Perceptual Justification Outside of Consciousness. In *Consciousness Inside and Out: Phenomenology, Neuroscience, and the Nature of Experience*, R. Brown (ed.), London: Springer, 137-145.

Berger, J., Nanay, B., & Quilty-Dunn, J. (2018). Unconscious perceptual justification. *Inquiry*, *61*(5–6), 569–589.

Berger, J. (2020). Perceptual consciousness plays no epistemic role. *Philosophical Issues*, 30(1), 7-23.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *The Behavioral and Brain Sciences*, *30*(5–6), 481–548.

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences*, *15*(12), 567–575.

Block, N. (2014). Rich conscious perception outside focal attention. *Trends in Cognitive Sciences*, *18*(9), 445–447.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071.

Boumans, M. (1999). Representation and stability in testing and measuring rational expectations. *Journal of Economic Methodology*, *6*(3), 381–402.

Boumans, M., & Morgan, M. S. (2001). Ceteris paribus conditions: materiality and the application of economic theories. *Journal of Economic Methodology*, *8*(1), 11–26.

Breitmeyer, B. G., Kafaligönül, H., Öğmen, H., Mardon, L., Todd, S., & Ziegler, R. (2006). Meta- and paracontrast reveal differences between contour- and brightness-processing mechanisms. *Vision Research*, *46*(17), 2645–2658.

Breitmeyer, B. G. (2014). *The visual (un)conscious and its (dis)contents*. Oxford: Oxford University Press.

Brogaard, B. (2015). Type 2 blindsight and the nature of visual experience. *Consciousness and Cognition*, *32*, 92–103.

Burge, T. (2010). *Origins of Objectivity* Oxford: Oxford University Press.

Byrne, A. (2016). The epistemic significance of experience. *Philosophical Studies*, *173*(4), 947–967.

Campbell, J. (2002). *Reference and consciousness*. Oxford: Oxford University Press.

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage*, *73*, 80–94.

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

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the Bandwidth of Perceptual Experience ? *Trends in Cognitive Sciences*, *20*(5), 324–335.

Drayson, Z. (2012). The uses and abuses of the personal / subpersonal distinction. *Philosophical Perspectives*, *26*(1), 1–18.

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology*, *54*(3), 304–313.

De Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a common currency between vision and audition. *PLoS ONE*, *11*(1).

Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J. F., Poline, J. B., & Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, *4*(7), 752–758.

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, *132*, 2531–2540.

Deroy, O., Spence, C., & Noppeney, U. (2016). Metacognition in multisensory perception. *Trends in Cognitive Sciences*, *20*(10), 736–747.

Dienes, Z., & 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.

Elgin, C. (2017). *True Enough*. Cambridge, MA: MIT Press.

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

Faivre, N., Arzi, A., Lunghi, C., & Salomon, R. (2017). Consciousness is more than meets the eye: a call for a multisensory study of subjective experience. *Neuroscience of Consciousness*, *3*(1), 1–8.

Feigl, H. (1958) The 'mental' and the 'physical'. *Minnesota Studies in the Philosophy of Science,* 2 (2), pp. 370–497.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1338–1349.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*(443), 1–9.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, *137*(10), 2811–2822.

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-Specific Disruption of Perceptual Confidence. *Psychological Science*, *26*(1), 89–98.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114.

Fleming, S. M. (2019). Awareness as inference in a higher-order state space. *Neuroscience of Consciousness*, *6*(1), 1–9.

Fleming, S. M. (2020). *Know Thyself: The Science of Self-Awareness*. Basic Books.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, 87(3):564–567.

Gajdos, T., Fleming, S. M., Saez Garcia, M., Weindel, G., & Davranche, K. (2019). Revealing subthreshold motor contributions to perceptual confidence. *Neuroscience of Consciousness*, *2019*(1), 1–8.

Gallagher, R., Suddendorf, T., & Arnold, D. (2018). Confidence as a diagnostic tool for perceptual aftereffects. *Scientific Reports*, (April), 270280.

Gallagher, R. M., Suddendorf, T., & Arnold, D. H. (2021). The implied motion aftereffect changes decisions, but not confidence. *Attention, Perception, and Psychophysics*, *83*(8), 3047–3055.

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

Gelbard-Sagiv, H., Faivre, N., Mudrik, L., & Koch, C. (2016). Low-level awareness accompanies "unconscious" high-level processing during continuous flash suppression. *Journal of Vision*, *16*(1), 1–16.

Green, D. and Swets, S. (1966). *Signal Detection Theory and Psychophysics.* Wiley, New York.

Grove, P. M., Ashton, J., Kawachi, Y., & Sakurai, K. (2012). Auditory transients do not affect visual sensitivity in discriminating between objective streaming and bouncing events. *Journal of Vision*, 12(8), Article 5.

Guggenmos, M. (2021). Measuring metacognitive performance: type 1 performance dependence and test-retest reliability. *Neuroscience of Consciousness*, *2021*(1), 1–14.

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences*, *9*, 1–66.

Hong, S. W., & Blake, R. (2009). Interocular suppression differentially affects achromatic and chromatic mechanisms. *Attention, Perception, and Psychophysics*, *71*(2), 403–411.

Hsieh, P. J., Colas, J. T., & Kanwisher, N. (2011). Pop-out without awareness: Unseen feature singletons capture attention only when top-down attention is available. *Psychological Science*, *22*(9), 1220–1226.

Hubel, D. H., & Wiesel, T. N. (1977). Ferrier Lecture: Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society B*, *198*, 1–59.

Huemer, M. (2001). *Skepticism and the Veil of Perception.*Lanham, MD: Rowman and Littlefield.

Huemer, M. (2006). Phenomenal Conservatism and the Internalist Intuition. *American Philosophical Quarterly*, 43(2),147‑158.

Irvine, E. (2012a). *Consciousness as a Scientific Concept.* Springer.

Irvine, E. (2012b). Old problems with new measures in the science of consciousness. *British Journal for the Philosophy of Science*, *63*(3), 627–648.

Irvine, E. (2013). Measures of consciousness.*Philosophy Compass*, *8*(3), 285–297.

Jenkin, Z. (2020). The epistemic role of core cognition.*Philosophical Review*, *129*(2), 251–298.

Jimenez, M., Villalba-García, C., Luna, D., Hinojosa, J. A., & Montoro, P. R. (2019). The nature of visual awareness at stimulus energy and feature levels: A backward masking study. *Attention, Perception, and Psychophysics*, *81*(6), 1926–1943.

Jin, M., & Glickfeld, L. L. (2019). Contribution of sensory encoding to measured bias. *Journal of Neuroscience*, *39*(26), 5115–5127.

Johnston, M. (2006). Better than mere knowledge? The function of sensory awareness. In T. S. Gendler & J. Hawthorne (Eds.), *Perceptual experience.* Oxford: Oxford University Press.

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

Kellij, S., Fahrenfort, J., Lau, H., Peters, M. A. K., & Odegaard, B. (2021). An investigation of how relative precision of target encoding influences metacognitive performance. *Attention, Perception, and Psychophysics*, *83*(1), 512–524.

Kentridge, R. W., Nijboer, T. C. W., & Heywood, C. A. (2008). Attended but unseen: Visual attention is not sufficient for visual awareness. *Neuropsychologia*, *46*(3), 864–869.

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the RoyalSociety B: Biological Sciences*, *367*(1594), 1322–1337.

Kiani, R., & Shadlen, M. N. (2009). Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science*, 324(5928), 759–64.

Kim, C., & Chong, S. C. (2021). Partial awareness can be induced by independent cognitive access to different spatial frequencies. *Cognition*, *212*(March), 104692.

King, J.-R., Pescetelli, N., & Dehaene, S. (2016). Brain Mechanisms Underlying the Brief Maintenance of Seen and Unseen Sensory Information. *Neuron*, *92*(5), 1122–1134.

Knotts, J. D., Odegaard, B., Lau, H., & Rosenthal, D. (2019). Subjective inflation: phenomenology's get-rich-quick scheme. *Current Opinion in Psychology*, *29*, 49–55.

Knotts, J. D., Michel, M., & Odegaard, B. (2020). Defending subjective inflation: an inference to the best explanation. *Neuroscience of Consciousness*, *2020*(1), 1–7.

Koivisto, M., & Neuvonen, S. (2020). Masked blindsight in normal observers: Measuring subjective and objective responses to two features of each stimulus. *Consciousness and Cognition*, *81*, 1–21.

Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception, and Psychophysics*, *77*(4), 1295–1306.

Koster, N., Mattler, U., & Albrecht, T. (2020). Visual experience forms a multidimensional pattern that is not reducible to a single measure: Evidence from metacontrast masking. *Journal of Vision*, *20*(3), 1–27.

Kunimoto, C., Miller, J. and Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses, *Consciousness and Cognition*, 10(3), 294–340.

LeDoux, J. E., Michel, M., & Lau, H. (2020). A little history goes a long way toward understanding why we study consciousness the way we do today. *Proceedings of the National Academy of Sciences of the United States of America*

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, 8(12):750–751.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores* Reading, MA: Addison-Wesley

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

Lycan, W. G. (2001). A simple argument for a higher-order representation theory of consciousness. *Analysis*, *61*(1), 3–4.

Macmillan, N. A. and Creelman, C. D. (2005). *Detection Theory: A User's Guide*. Taylor & Francis.

Macpherson, F. (2015). The structure of experience, the nature of the visual, and type 2 blindsight. *Consciousness and Cognition*, *32*, 104–128.

Malach, R. (2022). The Role of the Prefrontal Cortex in Conscious Perception: The Localist Perspective. *Journal of Consciousness Studies* 29 (7-8):93-114.

Mamassian, P. (2020). Confidence Forced-Choice and Other Metaperceptual Tasks. *Perception*, *49*(6), 616–635.

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

Maniscalco, B., & Lau, H. (2014). Signal Detection Theory Analysis of Type 1 and Type 2 Data: Meta-d', Response-Specific Meta-d', and the Unequal Variance SDT Model. In S. M. Fleming & C. D. Frith (Eds.), *The Cognitive Neuroscience of Metacognition* (pp. 25–66).

Markus, K. A., & Borsboom, D. (2013). *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. Routledge.

Mashour, G. A., Roelfsema, P., Changeux, J., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776–798.

Mazor, M., Moran, R., & Fleming, S. M. (2021). Stage 2 registered report: metacognitive asymmetries in visual perception. *Neuroscience of Consciousness*, *2021*(1), 1–15.

Mazor, M., & Fleming, S. M. (2020). Distinguishing Absence of Awareness from Awareness of Absence. *Philosophy and the Mind Sciences*, *1*(II), 1–13.

Mazzi, C., Bagattini, C., & Savazzi, S. (2016). Blind-sight vs. degraded-sight: Different measures tell a different story. *Frontiers in Psychology*, *7*(901), 1–11.

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., de Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(5), 1897–1906.

Meuwese, J. D. I., van Loon, A. M., Lamme, V. A. F., & Fahrenfort, J. J. (2014). The subjective experience of object recognition: Comparing metacognition for object detection and object categorization. *Attention, Perception, and Psychophysics*, *76*(4), 1057–1068.

Michel, M., & Morales, J. (2019). Minority Reports: Consciousness and the Prefrontal Cortex. *Mind & Language*, *35*(4), 493–513.

Michel, M. (2019). The Mismeasure of Consciousness: A problem of coordination for the Perceptual Awareness Scale. Philosophy of Science, 86(5), 1239–1249.

Michel, M. (2020). Consciousness Science Underdetermined: A short history of endless debates. *Ergo*, *6*(28), 771–809.

Michel, M., & Peters, M. (2020). Confirmation bias without rhyme or reason. *Synthese*, *199*, 2757–2772.

Michel, M. (2021). Calibration in Consciousness Science. *Erkenntnis*.

Michel, M., & Doerig, A. (2021). A new empirical challenge for local theories of consciousness. *Mind & Language*.

Michel, M., & Lau, H. (2021). Is Blindsight Possible Under Signal Detection Theory ? Comment on Phillips (2021). *Psychological Review*, *128*(3), 585–591.

Michel, M. (2022a). How (not) to underestimate unconscious perception. *Mind & Language*.

Michel, M. (2022b) Consciousness and the Prefrontal Cortex: A review. *Journal of Consciousness Studies.* 29(7-8), 115—157.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, 100(3):398–407.

Miyoshi, K., & Lau, H. (2020). A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychological Review*, *127*(5), 655–671.

Miyoshi, K., Sakamoto, Y., & Nishida, S. (Manuscript). On the assumptions behind metacognitive measurements: Implications for theory and practice. https://psyarxiv.com/3vxg5/.

Morales, J., Lau, H. & Fleming, S. M. (2018) Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal Cortex. *Journal of Neuroscience*, 38, 3534–3546.

Morales, J., & Lau, H. (n.d.). Confidence Tracks Consciousness. In J. Weisberg (Ed.), *Qualitative Consciousness: Themes from the Philosophy of David Rosenthal* Cambridge University Press.

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147.

Moreira, C. M., Rollwage, M., Kaduk, K., Wilke, M., & Kagan, I. (2018). Post-decision wagering after perceptual judgments reveals bi-directional certainty readouts. *Cognition*, *176*, 40–52.

Morgan, M. S. (2013). Nature's Experiments and Natural Experiments in the Social Sciences. *Philosophy of the Social Sciences*, *43*(3), 341–357.

Murphy, P. R., Robertson, I. H., Harty, S., & O'Connell, R. G. (2015). Neural evidence accumulation persists after choice to inform metacognitive judgments. *ELife*, *4*, 1–23.

Norman, E. and Price, M. C. (2015). Measuring consciousness with confidence ratings. In Overgaard, M., editor, *Behavioral Method in Consciousness Research*. Oxford: Oxford University Press.

Peirce, C. S., & Jastrow, J. (1884). On Small Differences of Sensations. *Memoirs of the National Academy of Sciences*, *3*, 75–83.

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

Persaud, N., & Cowey, A. (2008). Blindsight is unlike normal conscious vision: Evidence from an exclusion task. *Consciousness and Cognition*, *17*(3), 1050–1055.

Persaud, N., Davidson, M., Maniscalco, B., Mobbs, D., Passingham, R. E., Cowey, A., & Lau, H. (2011). Awareness-related activity in prefrontal and parietal cortices in blindsight reflects more than superior visual performance. *NeuroImage*, *58*(2), 605–611.

Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *ELife*, *4*(October), 1–30.

Peters, M. A. K., Ro, T., & Lau, H. (2016). Who's afraid of response bias? *Neuroscience of Consciousness*, *2016*(1), 1–16.

Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., Halgren, E., Lau, H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, *1*(7), 0139.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, 4(4), 171–212.

Phillips, I. (2016). Consciousness and Criterion: On Block's Case for Unconscious Seeing. *Philosophy and Phenomenological Research*, *93*(2), 419–451.

Phillips, I. (2018). Unconscious Perception Reconsidered. *Analytic Philosophy*, *59*(4), 471–514.

Phillips, I. (2021). Blindsight is qualitatively degraded conscious vision. *Psychological Review*, *128*(3), 558–584.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901.

Polat, U., & Sagi, D. (2007). The relationship between the subjective and objective aspects of visual filling-in. *Vision Research*, *47*(18), 2473–2481.

Pournaghdali, A., & Schwartz, B. L. (2020). Continuous flash suppression: Known and unknowns. *Psychonomic Bulletin and Review*, *27*, 1071–1103.

Quilty-Dunn, J. (2019). Perceptual Pluralism. *Noûs*, *54*(4), 807–838.

Rahnev, D., Nee, D. E., Riddle, J., Larson, A. S., & D'Esposito, M. (2016). Causal evidence for frontal cortex organization for perceptual decision making. *Proceedings of the National Academy of Sciences*, *113*(21), 6059–6064.

Rahnev, D., & Denison, R. N. (2018). Suboptimality in Perceptual Decision Making. *Behavioral and Brain Sciences*, *41*, 1–107.

Ramsoy, T. Z. and Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences*, 3(1):1–23.

Rausch, M., Müller, H. J., & Zehetleitner, M. (2015). Metacognitive sensitivity of subjective reports of decisional confidence and visual experience. *Consciousness and Cognition*, *35*, 192–205.

Rausch, M., & Zehetleitner, M. (2016). Visibility is not equivalent to confidence in a low contrast orientation discrimination task. *Frontiers in Psychology*, *7*(APR), 1–15.

Rausch, M., & Zehetleitner, M. (2017). Should metacognition be measured by logistic regression? *Consciousness and Cognition*, *49*, 291–312.

Rausch, M., Hellmann, S., & Zehetleiner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception & Psychophysics*, *80*(1), 134–154.

Rausch, M., Hellmann, S., & Zehetleitner, M. (2021). Modeling visibility judgments using models of decision confidence. *Attention, Perception, and Psychophysics*, *83*(8), 3311–3336.

Reingold, E. M. and Merikle, P. M. (1990). On the inter-relatedness of theory and measurement in the study of unconscious processes.*Mind & Language*, 5(1), 9–28.

Rolfs, M., Dambacher, M., & Cavanagh, P. (2013). Visual adaptation of the perception of causality. *Current Biology*, 23(3), 250–254.

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.

Rosenthal, O., Shimojo, S., & Shams, L. (2009). Sound-induced flash illusion is resistant to feedback training. *Brain Topography*, 21(3–4), 185–192.

Rosenthal, D. (2019). Consciousness and confidence. *Neuropsychologia*, *128*, 255–265.

Sackur, J. (2013). Two dimensions of visibility revealed by multidimensional scaling of metacontrast. *Cognition*, *126*, 173–180.

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

Samaha, J., & Denison, R. (2022). The positive evidence bias in perceptual confidence is unlikely post-decisional. *Neuroscience of Consciousness*, *2022*(1).

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.

Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., & Kepecs, A. (2021). Striatal dopamine mediates hallucination-like perception in mice. *Science*, *372*(6537).

Schmidt, T., Miksch, S., Bulganin, L., Jäger, G., Lossin, F., Jochum, J., & Kohl, P. (2010). Response priming driven by local contrast, not subjective brightness. *Attention, Perception, & Psychophysics*, *72*(6), 1556–1568.

Schmidt, T. (2015). Invisible stimuli, implicit thresholds: Why invisibility judgments cannot be interpreted in isolation. *Advances in Cognitive Psychology*, *11*(2), 31–41.

Sergent, C. & Dehaene, S. (2004). Is Consciousness a Gradual Phenomenon? *Psychological Science*, 15(11):720–728.

Shady, S., MacLeod, D. I. A., & Fisher, H. S. (2004). Adaptation from invisible flicker. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(14), 5170–5173.

Shanks, D. R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin and Review*, *24*(3), 752–775.

Shekhar, M., & Rahnev, D. (2018). Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *Journal of Neuroscience*, *38*(22), 5078–5087.

Shekhar, M., & Rahnev, D. (2020). Sources of Metacognitive Inefficiency. *Trends in Cognitive Sciences*, *25*(1), 12–23.

Shekhar, M., & Rahnev, D. (2021). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review*, *128*(1), 45–70.

Shepherd, J., & Mylopoulos, M. (2021). Unconscious perception and central coordinating agency. *Philosophical Studies*, *178*, 3869–3893.

Siedlecka, M., Hobot, J., Skóra, Z., Paulewicz, B., Timmermans, B., & Wierzchoń, M. (2019). Motor response influences perceptual awareness judgements. *Consciousness and Cognition*, *75*, 102804.

Siedlecka, M., Wereszczyński, M., Paulewicz, B., & Wierzchoń, M. (2020). Visual awareness judgments are sensitive to accuracy feedback in stimulus discrimination tasks. *Consciousness and Cognition*, *86*(September).

Siegel, S. (2017). *The Rationality of Perception*. Oxford: Oxford University Press.

Silins, N. (2011). Seeing through the "veil of perception." *Mind*, *120*(478), 329–367.

Skóra, Z., Ciupińska, K., Del Pin, S. H., Overgaard, M., & Wierzchoń, M. (2021). Investigating the validity of the Perceptual Awareness Scale – The effect of task-related difficulty on subjective rating. *Consciousness and Cognition*, *95*(March).

Smith, J. D., Couchman, J. J., & Beran, M. J. (2014). Animal Metacognition: A Tale of Two Comparative Psychologies. *Journal of Comparative Psychology*, *128*(2), 115–131.

Smithies, D. (2019). *The Epistemic Role of Consciousness*. Oxford: Oxford University Press.

Snodgrass, M., Bernat, E., Shevrin, H. (2004) Unconscious perception: a model-based approach to method and evidence. *Perception & Psychophysics*, 66, 846–867.

Spener, M. (2015). Calibrating introspection. *Philosophical Issues*, *25*(1), 300–321.

Spener, M. (2020). Consciousness, introspection, and subjective measures. In U. Kriegel (Ed.), *Handbook of the Philosophy of Consciousness*. Oxford University Press.

Staley, K. W. (2004). Robust Evidence and Secure Evidence Claims. *Philosophy of Science*, *71*(4), 467–488.

Staley, K. W. (2020). Securing the Empirical Value of Measurement Results. *The British Journal for the Philosophy of Science*, *71*(1), 1–34.

Stevens, S. S. (1946). On the theory of scales of measurement. Science, 103, 677–680.

Stevens, S. S. (1951). *Handbook of experimental psychology.* New York: Wiley.

Stober, R. S., Brussel, E. M., & Komoda, M. K. (1978). Differential effects of metacontrast on target brightness and clarity. *Bulletin of the PsychonomicSociety*, 12, 433–436.

Swets, J. A. (1973). The relative operating characteristic in psychology: A technique for isolating effects of response bias finds wide use in the study of perception and cognition. *Science*, 182(4116), 990–1000.

Swets, J. A. (1986a). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, 99(2), 181–198.

Swets, J. A. (1986b). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin*, 99(1), 100–117.

Szczepanowski, R., Traczyk, J., Wierzchoń, M., & Cleeremans, A. (2013). The perception of visual emotion: Comparing different measures of awareness. *Consciousness and Cognition*, *22*(1), 212–220.

Taylor, H. (2020). Fuzziness in the Mind: Can Perception be Unconscious? *Philosophy and Phenomenological Research*, *101*(2), 383–398.

Thompson, W. A., & Singh, J. (1967). The use of limit theorems in paired comparison model building. *Psychometrika*, *32*(3), 255–264.

Timmermans, B. and Cleeremans, A. (2015). How can we measure awareness ? An overview of current methods. In Overgaard, M., editor, *Behavioral methods in consciousness science.* Oxford University Press.

Trübutschek, D., Marti, S., Ueberschär, H., & Dehaene, S. (2019). Probing the limits of activity-silent non-conscious working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(28), 14358–14367.

van Fraassen, B. C. (2008). *Scientific Representation: Paradoxes of Perspective* Oxford: Oxford University Press.

van Vugt, B., Dagnino, B., Vartak, D., Safaai, H., Panzeri, S., Dehaene, S., & Roelfsema, P. R. (2018). The Threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science*, *360*, 537–542.

Vlassova, A., Donkin, C., & Pearson, J. (2014). Unconscious information changes decision accuracy but not confidence. *Proceedings of the National Academyof Sciences*, *111*(45), 16214–16218.

Vorberg, D., Mattler, U., Heinecke, A., Schmidt, T., & Schwarzbach, J. (2003). Different time courses for visual perception and action priming. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(10), 6275–6280.

Vul, E., & MacLeod, D. I. A. (2006). Contingent aftereffects distinguish conscious and preconscious color processing. *Nature Neuroscience*, *9*(7), 873–874.

Weiskrantz, L. (2009). *Blindsight: A case study spanning 35 years and new developments.* Oxford: Oxford University Press.

Wierzchoń, M., Paulewicz, B., Asanowicz, D., Timmermans, B., & Cleeremans, A. (2014). Different subjective awareness measures demonstrate the influence of visual identification on perceptual awareness ratings. *Consciousness and Cognition*, *27*(1), 109–120.

Witt, J. K., Taylor, J. E. T., Sugovic, M., & Wixted, J. T. (2015). Signal detection measures cannot distinguish perceptual biases from response biases. *Perception*, *44*(3), 289–300.

Xue, K., Shekhar, M., & Rahnev, D. (2021). Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Consciousness and Cognition*, *95*, 1–8.

Zehetleitner, M., & Rausch, M. (2013). Being confident without seeing: What subjective measures of visual consciousness are about. *Attention, Perception, and Psychophysics*, *75*(7), 1406–1426.

Zhaoping, L. (2008). Attention capture by eye of origin singletons even without awareness: A hallmark of a bottom-up saliency map in the primary visual cortex. *Journal of Vision*, 8(5).

Zhaoping, L., & Xiao, Z. (2016). Without informative cues, little can be learned to discriminate eye of origin of visual inputs after multiple weeks of training.*Journal of Vision*, 16, 440.

Zhaoping, L. (2019). A new framework for understanding vision from the perspective of the primary visual cortex. *Current Opinion in Neurobiology*, *58*, 1–10.

Zhang, X., Zhaoping, L., Zhou, T., & Fang, F. (2012). Neural Activities in V1 Create a Bottom-Up Saliency Map. *Neuron*, *73*(1), 183–192.