Submitted to *Anthology of Neuroscience & Philosophy* (Eds. Felipe de Brigard & Walter Sinnott-Armstrong), MIT Press.

The Neural Substrates of Conscious Perception without Performance Confounds

Jorge Morales[a], Brian Odegaard[b] & Brian Maniscalco[c1]

[a] *Johns Hopkins University, Department of Psychological and Brain Sciences*
[b] *University of Florida, Department of Psychology*
[c] *University of California Riverside, Department of Bioengineering*

Abstract: To find the neural substrates of consciousness, researchers compare subjects' neural activity when they are aware of stimuli against neural activity when they are not aware. Ideally, to guarantee that the neural substrates of consciousness—and nothing but the neural substrates of consciousness—are isolated, the only difference between these two contrast conditions should be conscious awareness. Nevertheless, in practice, it is quite challenging to eliminate confounds and irrelevant differences between conscious and unconscious conditions. In particular, there is an often-neglected confound that is crucial to eliminate from neuroimaging studies: task performance. Unless subjects' task performance is matched (and hence perceptual signal processing is matched), researchers risk finding the neural correlates of perception, rather than conscious perception. Here, we discuss the theoretical motivations for the performance matching framework and review empirical demonstrations of, and theoretical inferences derived from, obtaining differences in consciousness while controlling for task performance. We summarize signal detection theoretic modeling frameworks that explain how it is that we can derive performance-matched differences in consciousness without the effect being trivially driven by differences in criterion setting, and also provide principles for designing experimental paradigms that yield performance-matched differences in awareness. Finally, we address potential technical and theoretical issues that stem from matching performance across conditions of awareness, and we introduce the notion of "triangulation" for designing comprehensive experimental sets that can better reveal the neural substrates of consciousness.

1. Introduction

Imagine you are driving with your friend at night on a poor-lit road. Both of you have 20/20 vision, are used to driving at night, and are attentively looking at the road to ensure there are no surprises. Suddenly, your friend yelps: "Watch out!"—there is a large branch in the middle of the lane. You avoid it just in time, but only thanks to your friend's warning: you

[1] All authors contributed equally to this work.

had not noticed the branch before you were alerted to it. How can this be? How could, under similar perceptual circumstances, your friend experience the obstacle while you completely miss it? One plausible explanation is that your friend consciously saw the branch while you did not. At the crucial moment, your visual experience of the road did not include any debris—you were unaware of it.

This example illustrates important aspects of how philosophers and neuroscientists think about consciousness, which is commonly characterized as "what it is like" to be in a particular mental state (Nagel, 1974). For example, there is something that it is like to see the branch while, presumably, there is nothing it is like to be a camera that records the visual properties of the road. This notion of consciousness can be extended beyond visual experiences to include other sensory modalities (e.g. auditory or gustatory), feelings and emotions, pains, and perhaps even the conscious experience of having thoughts and desires.

This subjective feeling of what it is like to be conscious of a particular content (e.g. the branch on the road) is referred to as *phenomenal consciousness*. In contrast, *access consciousness* describes the functional aspects of being consciously aware of contents as they become accessible to cognitive systems such as working memory, reasoning, categorization, planning, decision-making, and more generally, rational control of action (Block, 1995, 2005, 2007).

Were you *phenomenally* aware of the branch but failed to *access* the conscious representation of it, such that your voluntary motor control mechanisms did not steer the wheel appropriately? Perhaps your conscious experience was rich and included the content <branch> but it was not accessible by your categorization and decision-making systems—which are supposed to have a more limited capacity and, at least in principle, they are distinct and independent from your phenomenal consciousness (Block, 2005, 2007, 2011; Lamme, 2010). Alternatively, perhaps your phenomenal experience of the road lacked any branch altogether; perhaps it did not include the content <branch>: there was no phenomenally conscious branch that your cognitive mechanisms failed to access (Phillips, 2011). Of course, it might well be the case that this sensible conceptual distinction does not reflect how things are split up in the mind and brain. Perhaps there is no phenomenal consciousness without access consciousness (Cohen, Cavanagh, Chun, & Nakayama, 2012; Cohen & Dennett, 2011; Stanislas Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006) or perhaps access consciousness capacity is not limited with respect to phenomenal consciousness (Gross & Flombaum, 2017). Even if these two types of consciousness are distinct in principle, it could be impossible to know what phenomenal experiences you are in if you cannot access them (Kouider, de Gardelle, Sackur, & Dupoux, 2010; Kouider, Sackur, & Gardelle, 2012).

Orthogonal to the phenomenal and access distinction, different things could be meant when we talk about consciousness (Rosenthal, 1993). We could mean transitive or content-consciousness, namely, when one is conscious of a particular content (e.g. being conscious of the branch); state-consciousness, namely, when a mental state itself is conscious (e.g. the conscious experience of seeing the branch in contrast to perceptually processing the branch albeit unconsciously); and creature-consciousness, namely, the overall conscious state of someone as an individual (e.g. someone awake compared to someone asleep, anaesthetized, or in a coma (Bayne et al., 2016).

The Scientific Study of Consciousness

Can we know, from a scientific point of view, what explains the difference in conscious contents between you and your friend? Theoretical and practical concerns may cause one to question the possibility of a scientific study of consciousness. From a purely theoretical standpoint, many philosophers and scientists share the intuition that studying access consciousness in general, and perceptual processing in particular, is "easy". That is, while understanding how we perceive the environment is challenging, understanding access consciousness and perceptual processing does not seem to pose a distinct theoretical challenge compared to other psychological and brain phenomena we study: perceptual and decision-making mechanisms compute information, and that is something that, at least in principle, we know how to study. In contrast, understanding phenomenal consciousness is sometimes considered to be "hard" (Chalmers, 1996). The idea is that even if we found what the neural correlates of conscious experiences are, these would still fail to explain why those biophysical processes give rise to those subjective experiences. This so-called hard problem of consciousness has garnered much attention in the last 25 years; however, not everyone shares the intuition that we should be troubled by the alleged irreducibility of consciousness (Bickle, 2008; Godfrey-Smith, 2008). The metaphysical assumptions of the problem can be rejected, as they involve a notion of deductive explanation that is too stringent (Taylor, 2016). Furthermore, phenomenal consciousness is supported by brain activity: understanding the neural substrates of consciousness should be within the purview of scientific research.

To study consciousness scientifically, researchers aim to create conditions that probe the thresholds of awareness, where stimuli are processed at an intermediate level of efficacy that yields graded levels of awareness ranging from complete unconsciousness to clear, full-blown awareness. These conditions may be achieved by, for example, presenting a mask right before or after the stimulus (forward/backward masking); presenting distinct images to each eye effectively yielding one of them invisible (binocular rivalry and continuous flash suppression); degrading the contrast or the presentation duration of the stimulus; using constant stimulation that, however, can be perceived in different ways (bistable figures); or disrupting visual processing with transcranial magnetic stimulation (TMS). Thus, these conditions allow scientists to contrast subjects' experiencing something against experiencing nothing (Fleming, 2019), i.e. *detection* (e.g. a branch versus nothing); or they can contrast experiencing *this* compared to experiencing *that*, i.e. *discrimination* (e.g. a branch versus a snake). Importantly, these contrasts can be characterized in an all or nothing fashion, or they can take into account relative levels of awareness, too. For example, you could be either aware or unaware of the branch, but you could also be less aware of the branch than your friend, or more aware now than you were before your friend yelped.

When searching for the neural substrates of consciousness, scientists look for the minimally jointly sufficient neural events required for having a conscious experience (Chalmers, 2000). To find these substrates, they compare the neural activity of subjects when they are (more) aware of stimuli against neural activity when they are not (or less) aware of them. When subtracting neural activity of the unconscious states from the conscious ones, the remaining activity should reveal the unique neural processes that support consciousness. Besides this kind of subtraction, scientists can also compare patterns of activity or connectivity profiles across conditions. Ideally, to guarantee that the neural substrates of consciousness—and nothing but the neural substrates of consciousness—are isolated, the only difference between these two contrast conditions should be phenomenal consciousness. For instance, the story at the beginning of the chapter would not be so surprising if you did not have 20/20 vision, if you were not paying attention, or if your friend had much more experience driving at night than you. Translating this scenario to the lab, this means that we need to ensure that the perceptual, attentional, and cognitive demands of a task, as well as the subjects' performance in it, are matched when subjects are aware and unaware. Then, and only then, we can expect to learn what the neural substrates of consciousness are.

Nevertheless, in practice, it is quite challenging to eliminate confounds and irrelevant differences between conscious and unconscious conditions. In particular, there is an often-neglected confound that is, however, crucial to eliminate from neuroimaging studies: task performance.

Task Performance: A Confound in Neuroimaging Studies

Task performance is the objective effectiveness with which subjects succeed in achieving an experiment's goal. On the road, the goal is to detect debris; your friend is objectively more effective at this task than you: their task performance is better than yours. In the lab, consider a task that consists in identifying the shape of a stimulus that is presented on the screen on multiple trials. A straightforward way of measuring someone's task performance is by computing the percentage of correct responses they provide across all the trials. Thus, task performance is an important reflection of subjects' capacity to process the perceptual signal (which is required for succeeding at the task at hand). However, when task performance differs across conscious and unconscious conditions, behavioral, perceptual, and cognitive capacities can be expected to differ as well. Most of the time, unless experimenters actively make an effort to match subjects' performance, performance is higher in conscious trials than in unconscious trials. On the road, when your friend is conscious of the branch, they are also more likely to detect its presence, to discern its location, to identify it as a branch and not a snake, etc. Problematically, because variations in awareness typically are closely correlated with variations in task performance, a direct comparison of neural activity during conscious and unconscious conditions risks revealing differences in the neural substrates of *perception* (and other behavioral and cognitive capacities) rather than, or in addition to, the neural substrates of *consciousness*. Consequently, matching performance is crucial in neuroimaging studies that compare neural activity across awareness conditions.

In the following sections we discuss the benefits and challenges of matching performance in consciousness research. In Section 2, we discuss the difference between subjective and objective measures of consciousness, how they dissociate, and argue that consciousness research needs to focus on subjective measures while keeping objective performance constant. Then, in Section 3, we elaborate on the logic of considering task performance a confound in neuroimaging studies of consciousness. In Section 4, we discuss methods based on Signal Detection Theory and the design of stimuli specifically created to match performance and still obtain differences in awareness. In Section 5, we discuss potential technical and theoretical issues that stem from matching performance across conditions of awareness. Finally, in Section 6, we discuss future directions in consciousness research, and introduce the notion of "triangulation" for designing comprehensive experimental sets that can better reveal the neural substrates of consciousness.

A note on terminology: unless otherwise specified, by "consciousness" we will refer specifically to phenomenal consciousness of visual contents as revealed by subjective reports in detection and discrimination tasks. The context should make clear whether we are discussing cases of all or nothing consciousness, or cases of relative levels of awareness.

2. Subjective and Objective Measures of Consciousness

To analyze neural data, experimenters need to know when subjects are conscious of the stimuli they are presented with and when they are not. A straightforward way to achieve this is by asking subjects to report their subjective state, e.g. "I saw the branch" or "I did not see a branch." For obvious reasons, this kind of *subjective measure* is widely used. However, subjective measures have been criticized both in philosophy and neuroscience. From a behavioral standpoint, critics argue that introspective reports of consciousness are prone to mistakes, biases and response criterion effects (Irvine, 2012; Phillips, 2016; Schwitzgebel, 2011; Spener, forthcoming). Subjects could report more or less frequently that they saw a stimulus due to their response strategies and not due to a reliable introspective judgment of their actual conscious experiences. By using *objective measures* that assess subjects' ability to detect and discriminate stimuli independently of whether they take themselves to have seen them consciously or not, experimenters could bypass the problem of the response criterion and the fallibility of introspection. From a neuroscientific perspective, an additional concern is that by eliciting subjective reports of consciousness, we risk capturing the neural correlates of *the report* of consciousness, instead of *consciousness itself* (Tsuchiya, Wilke, Frässle, & Lamme, 2015). To address this potential issue, critics have suggested using *no-report paradigms* where subjects' conscious status can be inferred by some indirect means other than direct subjective reports.

In this section, we discuss—and reject—the use of objective measures; instead, we argue that objective and subjective measures can come apart: you may report to be subjectively unaware of a stimulus and yet your behavior demonstrates that you are objectively able to detect or discriminate it (and vice versa). In the next section, we address the neuroscientific objections against subjective reports and argue that task performance is a confound in neuroimaging studies of consciousness.

Objective Measures

To assess the objective performance of a subject during a visual task, one can compute the percentage of their correct responses. But percentage correct estimates do not disentangle perceptual sensitivity from response bias. A more sophisticated method is estimating subjects' *d'* (d prime), which is a measure of perceptual sensitivity that stems from Signal Detection Theory (Green & Swets, 1966; Macmillan & Creelman, 2005). Importantly, one can estimate subjects' objective perceptual capacity (i.e. their perceptual signal-to-noise ratio; e.g. their ability to discern whether a line is tilted left or right) *independently* from their response bias (e.g. their overall propensity for reporting "left tilt" or "right tilt"). According to proponents of objective measures of consciousness, subjects' awareness of a stimulus can be equated with their perceptual sensitivity. Thus, if subjects do not perform a perceptual task above chance levels (i.e. *d'*=0), one could assume that they did not see the stimuli consciously (Holender, 1986; Kouider & Dehaene, 2007).

Unfortunately, the use of objective measures ignores a fundamental aspect of consciousness—in fact, it ignores what makes it an interesting phenomenon in the first place: its subjective character. In normal scenarios, perceptual sensitivity *may* track consciousness. For example, objectively discriminating branches from a clear road might coincide with the subjective report of experiencing a branch and the subjective report of experiencing no debris, respectively. However, as we show below, objective and subjective measures can dissociate: one can perceptually discriminate stimuli without awareness, and one can enjoy conscious experiences without any perceptual sensitivity. During illusions or hallucinations, conscious experiences do not entail perceptual discrimination above chance—during a hallucination there is nothing to discriminate! Alternatively, above-chance discrimination does not entail consciousness. For instance, artificial systems can make successful discriminations of visual stimuli, but with the current state of technology it is unlikely they are conscious (Stanislas Dehaene, Lau, & Kouider, 2017). Moreover, blindsight patients deny being conscious of perfectly visible stimuli presented in a blind region of their visual field, and yet, they are able to detect or discriminate these otherwise invisible stimuli significantly above chance. If we made *d'* the measure of awareness, we would need to reject patients' subjective reports. Rather than ignoring subjective reports, we should value them as an important window to awareness, which is distinct and dissociable from objective performance.

Subjective Measures

Subjective reports can be obtained using a wide variety of procedures, such as reports of awareness (e.g. "seen" vs "not seen" or "seen" vs "guess", as in e.g. Lau & Passingham (2006)); reports on the visibility of the stimulus (e.g. from "clearly visible" to "not visible", as in e.g. Sergent & Dehaene (2004)); the method of adjustment or comparative judgments between two stimuli, which allows estimation of the point of subjective equality (PSE) (e.g. "this stimulus is more visible than this other one", as in e.g. Knotts, Lau, & Peters (2018)); reports of awareness using the Perceptual Awareness Scale (PAS) (0=no awareness, 1=brief glimpse, 2=almost clear awareness, 3=clear awareness; (Ramsøy & Overgaard, 2004)); confidence ratings (e.g. 1=not confident, 2=barely confident, 3=somewhat confident, 4=very confident, as in e.g. Maniscalco & Lau (2012) or post-decision wagering (e.g. high vs low wager of points or money, as in e.g. Persaud, McLeod, & Cowey (2007)).

Although there are important differences among these subjective methods, they all aim to probe the qualities of subjects' conscious experiences. The first four methods require subjects to introspect and report on the nature of their experiences. Even though confidence ratings are more indirect, they are very commonly used in consciousness research. When asked to provide confidence ratings, subjects are asked about their *subjective* impression regarding their *objective* performance in the task. Despite being less direct, confidence ratings can provide similar insights into a subject's conscious experience as those given by direct introspective reports, while also potentially offering some advantages (but see Rosenthal (2019)). Empirically, confidence ratings often correlate with reports of subjective awareness (Michel, 2019; Peters & Lau, 2015; Sandberg, Timmermans, Overgaard, & Cleeremans, 2010). This empirical correlation reflects the fact that one's confidence in a visual task is largely shaped by one's phenomenology. If one sees clearly what is on the screen, in general one should be more confident that one responded correctly about the stimulus presence/identity; alternatively, if one is not clearly aware of the stimulus, one should be less confident in the correctness of their response—it should feel more like guessing (see Rausch & Zehetleitner (2016)). One potential advantage of confidence ratings is that it might be easier for subjects to understand what is being asked from them when providing confidence ratings than when they are asked to introspect about the nature of their subjective experience. A second advantage is that confidence ratings are more interpretable than awareness reports for assessing subjects' metacognitive capacity which, however, can potentially offer a meaningful window into subjective conscious states.

Metacognition is the capacity to monitor and evaluate one's own cognitive processes (Flavell, 1979; Fleming, Dolan, & Frith, 2012; Proust, 2013). Confidence ratings can be viewed as metacognitive judgments about the likelihood that a given response in a task is correct. As a consequence, it is possible to compute "objective" measures of metacognitive performance from subjective confidence ratings by quantifying how well confidence correlates with accuracy. In particular, signal detection theory analyses can provide a response-bias-free measure of metacognitive sensitivity analogous to *d'*, termed meta-*d'* (Maniscalco & Lau, 2012, 2014). This "objective" and response-bias-free measure thus offers the tantalizing potential for having the best of both worlds when studying awareness: taking subjective report seriously (like subjective measures), while sidestepping thorny issues of response bias (like objective measures) (Kunimoto, Miller, & Pashler, 2001). However, it is possible for blindsight patients to have above-chance metacognitive performance in their blind field (Persaud et al., 2011), and conceptually it is possible to have chance-level metacognition about phenomenological experiences (e.g. due to hallucination; Shaver, Maniscalco, & Lau (2008)), suggesting that the presence or absence of metacognitive sensitivity cannot be taken as a hard and fast indicator of the presence or absence of phenomenology (Maniscalco & Lau, 2012). Nonetheless, measures of metacognitive sensitivity may have heuristic value in assessing levels of stimulus awareness, as presumably one's metacognitive sensitivity would tend to dwindle with reductions in phenomenological stimulus awareness. For instance, Persaud et al. 2011 showed that although their blindsight patient had above-chance metacognitive performance in the blind field, this was still lower than metacognitive performance in the normally sighted field, in spite of the fact that visual task performance in the two fields was matched by stimulus titration.

Objective and Subjective Measures Can Dissociate

The idea that subjective and objective measures of consciousness can dissociate, and that their dissociation represents a unique opportunity to isolate the neural basis of conscious awareness, is not new. More than 20 years ago, Weiskrantz and colleagues (Weiskrantz, Barbur, & Sahraie, 1995) suggested that "blindsight" patients offer a stunning demonstration of how subjectivity and objectivity differ (Lau, 2008). Blindsight occurs when patients have damage to primary visual cortex (V1). These patients can perform many perceptual tasks at above-chance levels and yet report no phenomenological experience associated with this ability. In some patients, performance in the blind part of the visual field is as high as that of the unimpaired field, and phenomenological experience can be found in one, but not the other. Thus, blindsight patients provide a critical proof-of-principle in demonstrating how subjective and objective measures can dissociate within a single individual.

It is beyond the scope of the current work to exhaustively review the literature demonstrating the many ways in which objective and subjective measures can dissociate in healthy and atypical populations, although the examples of matched performance / different awareness findings discussed below constitute one salient subset of such evidence. Importantly, the dissociability of objective and subjective measures entails not only that objective measures may be unreliable indicators of consciousness, but also that differences in objective performance associated with differences in awareness can pose as confounds that must be controlled for in isolating the cognitive and neural properties of consciousness.

3. The Importance of Matching Task Performance

Pre-conditions, Concurrent-processing, and Post-processing Effects

An important challenge faced when trying to isolate the neural bases of consciousness is the need to distinguish, on one hand, the neural substrates of consciousness proper, and on the other hand, the pre- and post-processing that enable and follow conscious experiences, respectively (Aru, Bachmann, Singer, & Melloni, 2012). Equally important is to distinguish the processing that occurs concurrently with conscious processes, but that is ultimately irrelevant for supporting them. As indicated above, the proper neural substrates of consciousness are only those that are jointly minimally sufficient for sustaining a conscious experience with a given content (Chalmers, 2000; Shoemaker, 1981). There is, however, a multitude of pre-, concurrent-, and post-processes that are not sufficient (or even necessary) for sustaining conscious experiences. Some of these might be necessary for perceptually processing the stimulus (albeit unconsciously). Perhaps they are even necessary for giving rise to the neural events that are in fact the basis of consciousness, without themselves being a neural correlate of consciousness. Crucially, these irrelevant processes need to be eliminated or matched across conscious and unconscious conditions.

Consider comparing the neural activity of someone with their eyes open and then closed. They are more likely to consciously see a stimulus with their eyes open than with their eyes closed. However, comparing their neural states in these two conditions would hardly reveal the neural correlates of consciousness: so many other things are different! This extreme case illustrates what happens in more subtle scenarios where there are differences in pre-, concurrent-, and post-processing. For instance, consider the general excitability of neuronal populations. Oscillating pre-stimulus brain activity can reliably predict whether a subsequent stimulus is perceived or not (Benwell et al., 2017; Mathewson, Gratton, Fabiani, Beck, & Ro, 2009; Samaha, Iemi, & Postle, 2017). When contrasting conscious and unconscious trials, these differences in neural activity are likely to be reflected in neuroimaging data (specifically, in the phase of pre-stimulus alpha oscillations obtained in electroencephalography—EEG). However, these enabling pre-stimulus oscillations are not the neural substrate of consciousness.

Consider now post-processing. Consciously experiencing a stimulus is likely to have ripple effects in subsequent neural processing that are either lacking or reduced during unconscious perception. Some of these might be cognitive consequences that are not associated with consciousness at all (Block, 2019). For example, sustained maintenance of information in working memory, access to long-term memory, verbal reports, or intentional behavior are examples of post-perceptual processing that could be more markedly revealed in neural activity during conscious trials compared to unconscious trials. This post-processing neural activity, however, is not the neural substrate of consciousness proper as it only happens *after* consciousness has already started taking place (and of course it can overlap with continuing conscious processing).

This concern has led some researchers to argue that we need to eliminate subjective reports from consciousness research altogether (Tsuchiya et al., 2015). They worry that requiring subjective reports might reveal just post-processing neural activity associated with access and report itself, but not consciousness. However, it is important to curb these specific worries about subjective reports and highlight an important constraint: processing unrelated to consciousness is problematic *when it is not matched across conditions.* As long as subjects have similar cognitive and reporting requirements across conscious and unconscious trials, subjective reports need not be a confound (Michel & Morales, 2019).

Concurrent-processing of the stimulus (e.g. perceptual processing independent from consciousness such as distinguishing signal from noise, feature extraction, categorization, etc.), which is fundamental for performing the task successfully, takes place alongside processes supporting consciousness. But those perceptual processes are not part of the neural basis of consciousness since, presumably, these are perceptual processes that are also present during unconscious perception.

One might wonder, is there any neural activity left over? One important lesson from thinking about the importance of matching background conditions and cognitive processes across conditions of awareness is that the neural activity that supports consciousness may indeed be quite subtle. For instance, it might only be detectable with highly sensitive neuroimaging methods such as single-cell recording, sophisticated statistical methods such as multivariate (rather than univariate) analyses, and in localized rather than brain-wide activity (Morales & Lau, forthcoming). So, when pre-, concurrent-, and post-processes are not matched across conditions, experimenters risk conflating them with the neural substrates of consciousness proper. Unfortunately, while these differences might be conceptually clear, in practice it can be challenging to distill all these types of neural activity (Giles, Lau, & Odegaard, 2016). Part of the difficulty is that there is no clear temporal differentiation between relevant and irrelevant types of neural activity for consciousness. Activity related to pre-conditions could continue after stimulus presentation when neural activity related to consciousness begins. Similarly, the consequences of conscious awareness could begin to manifest while subjects are still aware of the stimulus, effectively creating temporally overlapping neural activity pertaining to distinct processes. Naturally, concurrent processes are especially hard to disentangle from conscious-related processes. Moreover, nothing we know about neurobiology rules out *a priori* that some pre-, concurrent-, and post-processing recruit at least a subset of the same neuronal populations recruited by consciousness processes.

An effective way to eliminate, or at least reduce, these confounds is to match the testing conditions across conscious and unconscious trials. As long as the pre-conditions, concurrent-processes, and post-effects of consciousness are sufficiently similar across conscious and unconscious trials, one may not need to worry about distilling them from the neural data pertaining to consciousness proper. This is because there is a reasonable expectation that they will cancel each other out. Some of the dimensions along which tasks are often matched include type, duration and strength of stimulation, response demands (e.g. sensorimotor and cognitive requirements for report), and cognitive demands (e.g. attention, working memory load, task difficulty, cognitive control, etc.). However, an important, yet often neglected, dimension that experimenters should match across conscious and unconscious trials is task performance.

Performance Matching is Key

Matching subjects' performance in conscious and unconscious trials ensures that concurrent perceptual signal processing is comparable. This is important both in itself and because it helps matching other types of processing. For instance, similarity in perceptual processing increases the odds that pre- and post-processing neural activity is comparable. If one wants to find the neural basis of consciousness proper, and distinguish it from the objective capacity to perceive a stimulus, performance matching is required. But it is also important because it correlates with other cognitive capacities. Whereas task performance can be straightforwardly computed (e.g. percentage of correct trials or d'), it is hard to objectively quantify cognitive processes such as cognitive effort, working memory load (beyond number of items to be reported), and so on. But matching these cognitive demands is important. By making sure that task performance is the same across conditions, we ensure that cognitive effort, working memory load, and other cognitive demands are similar as well.

While conceptually matching performance is desirable, it is hard to achieve in practice and it is in all likelihood impossible to achieve without creating differences somewhere else (see Section 6). To make someone unaware of an otherwise visible stimulus, some change in the testing conditions needs to take place (Kim & Blake, 2005). These changes can be applied to the stimulus itself (e.g. decreasing stimulus strength or duration, adding a mask or changing the mask's duration), to the task (e.g. increasing task difficulty), or to participants themselves (e.g. distracting participants' attention, altering their brain states directly via TMS).

It is important to emphasize that the goal of performance matching is to match *perceptual signal processing;* in other words, perceivers' capacity to process the perceptual signal triggered in their visual system such that it can disentangle signal from noise and eventually create a perceptual representation of the stimulus. To illustrate this point, consider the following case. Imagine an experiment where subjects detect stimuli correctly more frequently when they are conscious of them than when they are not—i.e. an experiment where performance, and hence perceptual signal processing, are not matched across conscious and unconscious trials. To fix this, one could try to "artificially match" for performance a posteriori by only analyzing the neural data of "correct" trials, leaving out "incorrect" trials. This way, performance in the selected trials would be, by necessity, matched at 100% in both cases. But this artificial correction would not necessarily match the *perceptual signal processing capacity* and its supporting brain states across different awareness conditions. For instance, in unaware trials in which subjects reported correctly, they could have guessed without perceptually processing the stimulus. One could attempt more sophisticated corrections to "guesses" in unaware trials by taking into account subjects' guessing rate (Lamy, Salti, & Bar-Haim, 2009). But this approach is insufficient for matching the underlying perceptual capacity and the corresponding neural activity that drives correct trials in aware and unaware conditions (Morales, Chiang, & Lau, 2015). Thus, artificial matching should be avoided.

Performance Matching Reveals Neural Correlates of Consciousness in PFC, in Agreement with Higher-Order Theories

One seminal demonstration of performance matching comes in a metacontrast masking study by Lau & Passingham (2006). In their behavioral experiment, subjects were presented with a brief visual target and were required to discriminate its identity (either diamond or square) and indicate whether they consciously saw the target or not. Critically, a metacontrast mask was presented with varying stimulus onset asynchrony (SOA) after the visual target. Behavioral results showed that two distinct SOAs yielded similar levels of performance on the discrimination task, but different levels of awareness (the percentage of trials subjects reported seeing the stimulus). Functional magnetic resonance imaging (fMRI) revealed that while activations in many cortical areas distinguished performance levels in general (i.e. correct vs. incorrect trials), only dorsolateral prefrontal cortex (DLPFC) activity reflected differences in two SOA conditions with matched performance and different awareness.

Maniscalco & Lau (2016) replicated the behavioral effect and conducted a model comparison analysis to test the ability of various candidate theories to capture the data. They found that the data were best captured by models embodying principles of higher-order theories of consciousness (Brown, Lau, & LeDoux, 2019; Lau & Rosenthal, 2011), in which task performance is determined by "first-order" processing and conscious awareness is determined by subsequent "higher-order" processing that evaluates first-order processing.

Lau & Passingham's finding that performance-matched differences in awareness are associated with activity in DLPFC but not sensory cortices can be well accommodated by higher-order theory given the broad observation that various forms of first-order processing tend to occur in posterior sensory cortices, whereas higher-order processing is more localized to prefrontal cortex (Brown et al., 2019).

The special role of prefrontal cortex in supporting subjective awareness, independently of objective task performance, is supported by a number of other studies. Disruption of DLPFC function by TMS (Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010; Ruby, Maniscalco, & Peters, 2018) or concurrent task demands (Maniscalco & Lau, 2015) selectively impairs metacognitive sensitivity but not objective performance in perceptual tasks. Patients with anterior prefrontal cortex lesions exhibit selective impairment of metacognitive sensitivity on a perceptual task relative to temporal lobe patients and healthy controls, even when task performance is matched across groups (Fleming, Ryu, Golfinos, & Blackmon, 2014). In a blindsight patient, frontoparietal areas in the brain are more activated for stimulus perception in the healthy visual field than in the blind visual field, even when task performance across the fields is equated (Persaud et al., 2011). Metacognitive sensitivity and task performance dissociate over time as one continuously performs a demanding task without rest, and this dissociation can be accounted for by individual differences in grey matter volume in anterior prefrontal cortex (Maniscalco, McCurdy, Odegaard, & Lau, 2017). Higher prestimulus activity in the dorsal attention network is associated with lower confidence ratings but not altered task accuracy (Rahnev, Bahdo, de Lange, & Lau, 2012). Further examples of matched performance with different awareness are discussed in the next section.

4. Understanding and Designing Matched Performance, Different Awareness Stimuli with Signal Detection Theory

A general principle that has been employed to both explain and generate matched performance with different awareness data is that task performance depends on signal-to-noise ratio, whereas awareness often depends more so on absolute levels of perceptual evidence. For instance, imagine a simple signal detection theory model in which two stimulus classes, S1 and S2, generate normal distributions of perceptual evidence along a decision axis (Fig. 1A), such that the perceptual evidence elicited by presentation of a stimulus on a given trial is a random draw from the corresponding evidence distribution. Suppose that in condition A (Fig. 1A, top panel), the S1 and S2 distributions have means at the decision axis values -1 and $+1$, respectively, and standard deviations of 1. The subject responds "S2" if the evidence value e on the current trial exceeds 0, and endorses classification responses with high confidence if $e<-2$ or $e>2$ (corresponding to strong evidence for S1 or S2, respectively). Now suppose that in condition B (Fig. 1A, bottom panel), the S1 and S2 distributions have means of −2 and +2, and standard deviations of 2, but that the subject's decision rules for classifying and rating confidence remain the same. Conditions A and B then have identical task performance due to having identical signal-to-noise ratio; in both cases, the means of the evidence distributions are two standard deviations apart (i.e. d'=2), meaning it is equally difficult to infer whether a given perceptual sample originated from S1 or S2. However, confidence is higher in condition B, since in this case the absolute levels of perceptual evidence are more extreme and therefore more frequently exceed the criteria for high confidence. In this way, provided that the subject uses the same decision strategy across conditions[2], higher absolute levels of evidence will cause higher confidence even for matched signal-to-noise ratios.

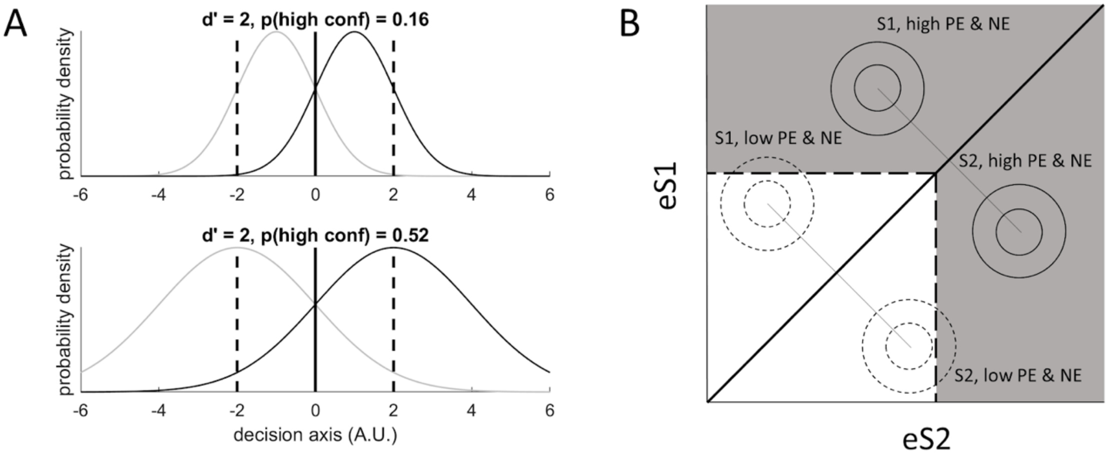

Figure 1. **Explaining matched performance, different awareness with one-dimensional and two-dimensional signal detection theory. (A)** In the one-dimensional case, consider a scenario where a subject needs to decide whether a given evidence sample on the decision axis was drawn from class S1 (gray distribution) or S2 (black) and rate confidence. The discrimination judgment depends on whether the given sample drawn is above or below a decision criterion, which in this example is set at 0 (solid line). If the sample is below 0, the subject selects S1, and if the sample is above 0, the subject selects S2. The rating of low or high confidence depends on where the sample falls with respect to the confidence criteria (dotted lines). In this example, samples greater than 2 or less than -2 yield high confidence ratings, while samples within this range yield low confidence. It follows that in the bottom panel, average confidence is higher due to higher evidence variance, in spite of task performance (signal-to-noise ratio, *d'*) being the same as in the top panel. **(B)** In the two-dimensional case, the two axes represent evidence for each stimulus class (eS1 and eS2). Circles represent bivariate normal distributions, and samples drawn from these distributions thereby contain

[2] Note that experimental designs in which conditions are randomly interleaved across trials can help ensure that decision strategy is constant across conditions, since human subjects have difficulty dynamically adjusting response criteria from trial to trial even when it would be ideal to do so (Brown & Steyvers, 2005; Gorea & Sagi, 2000).

evidence for both eS1 and eS2. Thus, the means of these distributions represent various positive evidence (PE) / negative evidence (NE) levels. Discriminating whether the stimulus is S1 or S2 involves evaluating whether the sample falls above or below the solid diagonal line. Confidence, however, involves evaluating the magnitude of the response-congruent evidence, which is shown by the confidence criteria (dashed lines) separating the white and gray regions. In this example, the high PE / NE stimuli (solid circles) have higher confidence than the low PE / NE stimuli (dotted circles) due to having more probability mass exceeding the confidence criteria, in spite of task performance (distance between the means of the distributions divided by standard deviation, *d'*) being the same.

This mechanism has successfully modeled performance-matched differences in awareness due to attentional manipulations (Rahnev et al., 2011), as well as simultaneous decreases in task performance and increases in confidence caused by TMS (Rahnev, Maniscalco, Luber, Lau, & Lisanby, 2012). It has also generated novel matched performance, different awareness findings by informing the experimental design of stimuli. For instance, experiments 1B and 2B of (Koizumi, Maniscalco, & Lau, 2015) used random dot motion stimuli in which a subset of dots moved coherently left or right, and the rest moved randomly. Across conditions, the fraction of coherently moving dots (i.e. signal-to-noise ratio) was the same, but the overall number of dots (i.e. absolute levels of perceptual evidence) differed. As expected, confidence was higher for stimuli with higher dot density, even though task performance was the same. (Samaha, Barrett, Sheldon, LaRocque, & Postle, 2016; Samaha, Switzky, & Postle, 2019) used oriented gratings in noise as stimuli and employed similar manipulations of the stimuli. Across conditions, the ratio of grating contrast to noise contrast was identical, but overall contrast of the composite stimulus differed, yielding higher confidence in the higher contrast stimuli despite equivalent performance.

A similar principle and accompanying method of stimulus construction comes from findings that confidence follows a *response-congruent evidence* rule. That is, confidence depends heavily on evidence congruent with the perceptual decision while downweighting or ignoring evidence that contradicts the perceptual decision (Zylberberg et al. 2012; Maniscalco et al. 2016; Peters et al. 2017). Exploiting this finding, experiments 1A and 2A of Koizumi et al. (2015) used stimuli with different levels of positive evidence (PE) and negative evidence (NE), where PE is evidence supporting the correct perceptual decision and NE is evidence supporting the incorrect decision. Specifically, they used oriented gratings embedded in noise, where a higher-contrast grating (PE) tilted left or right was superimposed with a lower-contrast grating (NE) tilted in the opposite direction, and the correct tilt response corresponded to the higher-contrast grating. By manipulating the contrasts of PE, NE, and noise, they created conditions where performance was similar but PE and NE levels differed. Crucially, since confidence depends on response-congruent evidence, confidence was higher in the conditions with higher PE and NE levels.

An illustration of the logic of capitalizing on the response-congruent evidence rule to create matched performance, different awareness stimuli by manipulating PE and NE levels is shown in Figure 1B. Following Maniscalco, Peters, & Lau (2016), we use a two-dimensional signal detection theory representation in which the two axes, eS1 and eS2, correspond to evidence for the two stimulus classes S1 and S2. Generalizing from the one-dimensional case (Figure 1A), we assume that each stimulus class generates a bivariate normal distribution, such that the perceptual evidence elicited by a stimulus on a given trial is a random draw of an (eS1, eS2) pair from the corresponding stimulus distribution. Circles in the plot represent contours of the distributions as 3D hills seen from above, similar to a topographic map. The mean of the distributions corresponds to PE and NE levels; for instance, an S1 stimulus with high PE and intermediate NE will have a high mean value along the eS1 dimension and an intermediate mean value along the eS2 dimension. Given evidence (eS1, eS2) on a given trial, the subject responds "S2" if eS2>eS1 (region of the plot below the solid diagonal line eS1=eS2), and "S1" otherwise. Crucially, the subject rates confidence by comparing the magnitude of response-congruent evidence to a criterion value (corresponding to the dashed horizontal and vertical lines), yielding high confidence for evidence pairs located in the shaded region of the plot.

In Figure 1B, we show stimulus distributions for two experimental conditions, one with low PE and NE, and one with high PE and NE. Task performance (*d'*) is determined by the distance between the distributions along the line connecting their means, divided by their common standard deviation. Thus, the low and high PE/NE conditions shown here have matched levels of *d'*. However, a greater proportion of the high PE/NE distributions lies within the shaded region of the plot than the low PE/NE distributions, thus yielding higher confidence. Note that this arrangement depends on the response-congruent evidence rule in order to yield differences in confidence; if confidence depended on the magnitude of the difference in evidence eS2−eS1, then the dashed confidence criterion lines would be 45 degrees (parallel to the solid perceptual decision criterion), and the proportion of the distributions lying in the shaded (high confidence) regions would be equivalent for the high and low PE/NE stimuli.

Notably, the PE/NE method of creating stimuli yielding matched performance and different awareness has the advantage that it allows for overall stimulus energy (e.g. contrast or dot density) to be matched across conditions, since increases in PE and NE energy can be offset by decreases in noise energy. By contrast, the signal/noise method requires there to be higher overall stimulus energy in the condition with higher awareness, thus posing an undesirable confound. On the other hand, PE/NE manipulations potentially induce response conflict in a way that signal/noise manipulations don't (by virtue of PE and NE priming opposing perceptual decisions / responses), which can also be undesirable.

Note that the mechanisms discussed in this section are not meant to be exhaustive explanations for all cases; it is possible that other kinds of mechanisms can also produce matched performance, different awareness data, such as the higher-order model of Maniscalco & Lau (2016) mentioned previously. Nonetheless, the methods discussed in this section are powerful insofar as they not only provide potential post-hoc explanations, but actually enable us to *design* stimuli that yield matched performance and different awareness using well-understood computational principles.

5. Theoretical Caveats and Nuances

We can summarize the logic of performance matching as follows: to precisely isolate subjective awareness of a stimulus from confounding factors, we should conduct experiments that satisfy the following criteria:

1. **Dissociable processing identified:** We have some notion of which sensory and perceptual processing of the stimulus is dissociable from awareness and thus needs to be controlled for when experimentally isolating awareness
2. **Dissociable processes matched:** We empirically confirm that the dissociable processing identified in (1) is matched across experimental conditions by demonstrating equal performance on a task that probes such processing
3. **Awareness differs:** Average subjective awareness of the stimulus differs across conditions

Here we highlight some nuances and potential difficulties in each of these criteria that should inform the way we conduct and interpret performance matching studies and the study of subjective awareness more broadly. In brief, the nuances explored for each criterion are:

1. **Uncertainty about dissociable processing:** There is some uncertainty about which perceptual processing is dissociable from awareness and which is not
2. **Multidimensionality of dissociable processing:** There are potentially many dimensions of stimulus processing that are dissociable from awareness other than the task probed in the experiment
3. **Absolute vs relative levels of awareness:** Interpreting a difference in awareness requires considering not just the relative difference in reported awareness across conditions, but also the absolute level of awareness within each condition

Uncertainty about Dissociable Processing

In order to argue that we should control for some aspect of perceptual processing *P* when studying awareness, we must have some prior reason for thinking that *P* is dissociable from awareness to begin with. For instance, we have strong reason to believe that forced-choice discrimination of simple stimulus features can proceed without awareness from blindsight patients (Weiskrantz, 1986). There is also evidence for above-chance forced-choice stimulus discrimination in healthy observers (e.g. Kouider, Dehaene, Jobert, & Le Bihan, 2007; Merikle, Smilek, & Eastwood, 2001; Snodgrass, Bernat, & Shevrin, 2004), although such findings are more contentious (Eriksen, 1960; Hannula, Simons, & Cohen, 2005; Lloyd, Abrahamyan, & Harris, 2013; Peters & Lau, 2015; Phillips, 2016). And of course, as reviewed above, there is ample evidence that awareness can differ across conditions with matched forced-choice discrimination performance.

Stances on what aspects of stimulus processing are dissociable from awareness, versus which are inseparable from or deeply intertwined with it, are influenced not just by evidence but also theory. For instance, some theoretical frameworks—such as higher order theories (Brown et al., 2019; Lau & Rosenthal, 2011) and some interpretations or implementations of signal detection theory (Maniscalco et al., 2016)—lend themselves naturally to viewing task performance and subjective reports of awareness as strongly dissociable, whereas other frameworks posit a tighter relationship in which cleanly separating task performance and awareness might not always be so straightforward. For instance, in Global Workspace theory (e.g. Baars, 2005; Baars, 1997; Dehaene, 2014), a content becomes conscious by virtue of entering a "global workspace," but also enjoys enhanced processing by virtue of being in the workspace, such that the enhanced processing of the content may not be completely separable from awareness of the content *per se*.

Importantly, these theoretical orientations affect not just predictions about what sorts of stimulus processing should be dissociable from awareness, but also interpretation of extant demonstrations of such dissociations. For instance, for a higher-order theorist, matched performance dissociations are straightforward demonstrations of the theoretically expected separability of task performance and awareness. By contrast, a global workspace theorist might hold that even though feature discrimination and stimulus awareness are partially dissociable, nonetheless awareness of a stimulus plays some direct participatory role in the full-blown kind of feature discrimination present in conditions of full stimulus awareness. (An instance of such a view is the model of Del Cul, Dehaene, Reyes, Bravo, & Slachevsky, 2009.) For such a theorist, matching discrimination performance when studying awareness might eliminate too much, removing the confounds of non-conscious contributions to feature discrimination while also masking the contributions of consciousness itself, leaving only some minimal difference in awareness that happens to be insufficient to manifest as a difference in task performance.

More generally, to whatever extent awareness directly participates in some aspect of perceptual processing *P*, that aspect of awareness must necessarily be masked by experimental procedures that match *P* across conditions. Demonstrating that awareness can differ to some extent when *P* is matched does not necessarily entail that awareness plays no part in *P* whatsoever; it only conclusively demonstrates that in some conditions, it is possible for the observed difference in awareness to fail to manifest as a difference in *P*.

There is thus a kind of circularity that poses some difficulty for different theoretical camps to agree on basic aspects of methodology in consciousness science: our theories should be constrained by empirical results, but the interpretation of those results and how they should refine our theories is itself theory-dependent. Continued advances in empirical findings and theoretical developments will presumably lead to increasing convergence on both theory and methodology, but achieving such convergence is nontrivial in the face of these issues.

Multidimensionality of Dissociable Processing

To this point, we have focused the discussion on matching task performance for the task being probed in the experimental design. For instance, if the task requires the subject to discriminate left vs right grating tilt and then report awareness, we would recommend to study awareness by comparing two experimental conditions where tilt discrimination performance is equal and yet average subjective report differs. However, it is of course the case that the subject performs many other perceptual operations ("tasks") that are not directly probed by such a design, e.g. detecting the presence of the grating, identifying the detected stimulus *as* an oriented grating, discerning the exact degree of its tilt (as opposed to making a binary left / right classification), etc.

We can therefore differentiate between *probed* task performance (performance on the task explicitly measured in the experiment, e.g. tilt discrimination) and *latent* task performance (performance on perceptual "tasks" that were not explicitly probed in the experiment but could have been, e.g. stimulus detection, object identification, etc.). The question then becomes whether matching *probed* task performance is sufficient for matching *latent* task performance. Presumably, as general quality of stimulus processing improves (e.g. due to stronger stimulus drive, improved attention, etc.), different dimensions of perceptual processing (detection, feature discrimination, identification, etc.) will all improve as well. The existence of such a correlation in perceptual performance across different dimensions of stimulus processing helps address concerns about possible confounds in latent task performance when probed task performance is matched. Yet, there is no general guarantee that matching performance on the probed task entails matching performance on all latent tasks. For instance, it can be readily demonstrated with a signal detection theory model that identical levels of performance for discriminating between stimuli A and B are compatible with different levels of performance for detecting A and B. For instance, by increasing the means and variances of the evidence distributions appropriately, *d'* can remain unchanged (as in Fig. 1A), but such increases in mean and variance will yield altered detection performance with respect to a stimulus-absent noise distribution with fixed mean and variance.

Of course, it is impossible in practice to probe all relevant kinds of perceptual processing in a single experiment, and so a pragmatic approach is just to match performance on a representative task (such as feature discrimination) and assume that this does an acceptable job of matching latent task performance. However, it is worth keeping in mind that the same logic that would lead us to worry about matching probed task performance should also lead us to worry about matching latent task performance. If latent task performance is not matched, then between-condition differences in behavior or neural activity could potentially be attributed to the difference in a "latent task" rather than the difference in awareness *per se*. Additionally, it is possible that in some situations we might have reasons to believe that some aspect of latent performance *is not* matched in spite of matched performance on the probed task, and such situations would require special care (e.g. caution in interpreting the results, or designing a new study that properly controls for the task performance in question).

Absolute Levels of Awareness

Not all differences in awareness are created equal. For instance, imagine an experiment where subjects use the perceptual awareness scale (PAS) (Ramsøy & Overgaard, 2004), a standardized scale for rating visual awareness with four levels: no awareness (PAS rating=0), brief glimpse (rating=1), almost clear awareness (rating = 2), and clear awareness (rating=3). A performance-matched difference in PAS levels of 0 (no reported awareness whatsoever) and 1 (the first hints of entry of the stimulus into awareness) would then indicate something very different from a difference in PAS levels of 2 (almost clear awareness) and 3 (clear awareness). In turn, this would have consequences for interpreting the performance-matched difference in awareness in terms of cognitive functions or neural mechanisms. An experiment achieving a performance-matched difference of PAS rating=0 vs 1 would allow inferences about what cognitive functions and neural mechanisms correspond to the transition of a stimulus representation from complete unconsciousness to the first faint entries into conscious awareness[3]. By contrast, an experiment achieving a performance-matched difference of PAS rating=2 vs 3 would not allow inferences about the functions and mechanisms of a representation's being conscious as such, but rather would be limited to

[3] For simplicity, here we bracket legitimate concerns about response biases that complicate taking such reports at face value.

inferences about the cognitive functions and neural mechanisms corresponding to increases in the relative intensity or clarity of contents that are already conscious. (For a related discussion, see Michel, 2019.)

Studies on awareness are often centrally interested in the cognitive functions and neural mechanisms of a stimulus representation being conscious as such. In principle, the ideal way to approach this research question from a performance-matching perspective would be to achieve performance-matching for a "completely unconscious" condition[4] (i.e. PAS=0) and a "somewhat conscious" condition (PAS > 0). In practice, performance matching studies to this point have typically compared conditions in which subjects report an intermediate level of stimulus awareness in both conditions (Koizumi et al., 2015; Lau & Passingham, 2006; Maniscalco & Lau, 2016; Samaha et al., 2016), making them ideally suited to investigating the relative degree of intensity or clarity of contents of awareness, rather than awareness *per se*.[5] It has furthermore been difficult to unambiguously demonstrate above-chance performance without awareness in healthy subjects (Eriksen, 1960; Hannula et al., 2005; Lloyd et al., 2013; Peters & Lau, 2015). Thus, using the performance-matching framework to study the cognitive functions and neural mechanisms of consciousness as such, as opposed to the functions and mechanisms of changes in intensity or clarity of contents that are already conscious, faces significant practical hurdles still in need of addressing in future work.

Another way in which the absolute levels of awareness in performance-matched conditions matter is in interpreting the potential role of awareness in supporting further cognitive functions. For instance, (Koizumi et al., 2015) used specially designed grating stimuli to yield performance-matched differences in confidence for discriminating grating tilt. Confidence was rated on a scale of 1 to 4. Across two levels of *d'* for tilt discrimination, mean confidence for the low and high confidence stimuli was about (low=2, high=2.3) and (low=2.3, high=2.5) respectively. Koizumi et al. then used the tilt of the performance-matched stimuli as cues in go/no-go and task set preparation tasks to probe the role of performance-matched differences in confidence on cognitive control. They found that higher confidence did not confer an advantage in either cognitive control task. As with any null effect, this finding needs to be interpreted with caution; failure to find an effect could be due to a true absence of an effect (Figure 2A), or failure to detect a true but weak effect (Figure 2B).

In addition to these possibilities, it is also possible that failure to find an effect reflects a ceiling effect (Figure 2C) or floor effect (Figure 2D) attributable to the absolute levels of confidence probed in this study. For instance, it is possible that increases in performance-matched confidence do increase cognitive control, but that this effect is most pronounced at

[4] Again, bracketing response bias concerns.

[5] Such concerns could be somewhat alleviated if it could be demonstrated that the difference in average awareness is driven strongly by different frequencies of "completely unconscious" or PAS=0 trials.

lower levels of confidence (confidence<2) and is already saturated for the levels of confidence probed in Koizumi et al. (confidence>2) (Figure 2C). Alternatively, it could be that performance-matched increases in confidence only manifest as increases in cognitive control for higher absolute levels of confidence than were probed in Koizumi et al. (Figure 2D).

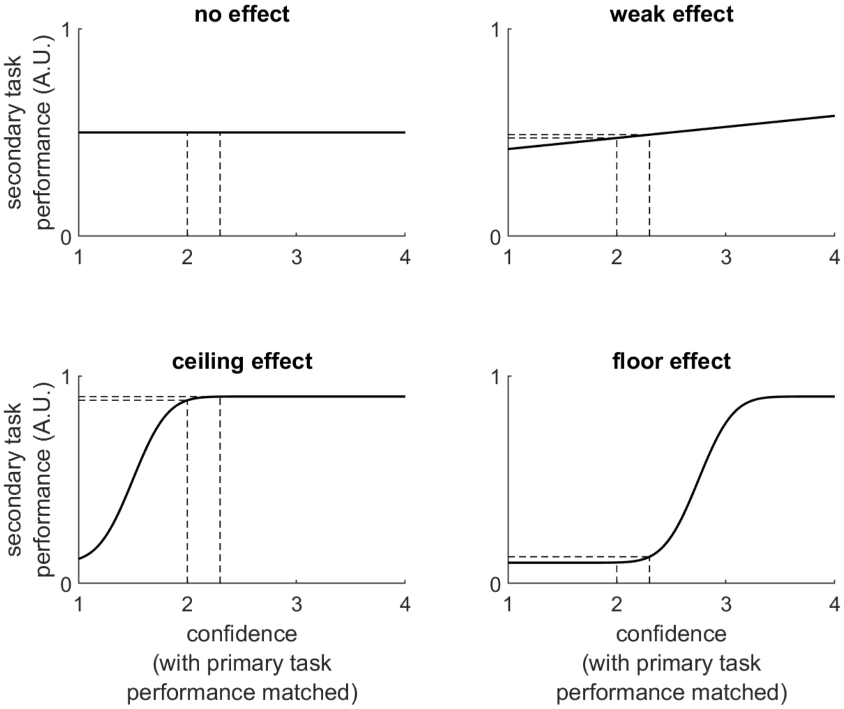

***Figure 2.* Possible ways in which performance-matched differences in awareness could fail to yield differences on a secondary task.** Plots show confidence on the x-axes, with the idealized assumption that these levels of confidence are all achieved at a constant level of primary task performance. Plotted on the y-axes are performance on a secondary task (such as the cognitive control task in Koizumi et al. 2015) using the matched performance, different awareness stimuli (e.g. using grating tilt to inform task set preparation for a separate, upcoming task). Vertical lines indicate two levels of performance-matched confidence probed in a hypothetical experiment, and horizontal lines show the corresponding difference in the secondary task under different scenarios. (A) *No effect*: performance-matched confidence does not affect the secondary task. (B) *Weak effect*: the influence of performance-matched confidence is small and therefore difficult to detect in an experiment. (C) *Ceiling effect*: performance-matched confidence does influence the secondary task, but the effect is saturated at the levels of confidence probed in the experiment. (D) *Floor effect*: performance-matched confidence does influence the secondary task, but the effect is stronger at higher levels of confidence than those probed in the experiment.

6. Future Directions: Triangulating on Consciousness

We have argued that task performance is a serious yet underappreciated confound in the neuroscientific study of consciousness. Yet, even if we can find conditions yielding different levels of stimulus awareness while task performance is matched—and even if we satisfactorily address the caveats and nuances discussed in the previous section—the unfortunate fact remains that *some* confound in the comparison between the "more conscious" and "less conscious" conditions must be present. Namely, there must be some difference between the conditions that causes awareness to differ, whether it is a difference in stimulus properties, attention, brain stimulation, or some other factor.

In practice, stimulus confounds are the type of confound most likely to be salient for performance matching studies. The "matched signal-to-noise ratio, different variance" method for designing performance-matched stimuli discussed previously (Fig. 1A) requires energy in the signal, noise, and overall stimulus to be larger in the "more conscious" condition. The "positive evidence / negative evidence" method (Fig. 1B) allows for overall stimulus energy to be matched, but only if energy in stimulus noise in the "conscious" condition is reduced to compensate for the increases in the energy of positive and negative evidence necessary to yield higher levels of awareness (Koizumi et al. 2015). These stimulus confounds are more severe than is typically encountered in more traditional consciousness experiments, where stimulus confounds are frequently minimal (e.g. differences in the temporal gap between stimulus and mask on the order of tens of milliseconds, as in Dehaene et al., 2001) or non-existent (e.g. a fixed stimulus repeatedly presented at threshold contrast so that it is sometimes consciously experienced and other times not, as in Baria, Maniscalco, & He, 2017).

Yet, of course, these studies invoking minimal or no stimulus confound suffer from drastic performance confounds. (A notable exception here is the metacontrast masking paradigm employed in Lau & Passingham (2006) and Maniscalco & Lau (2016), which can achieve performance matching with a difference in stimulus-mask onset asynchrony on the order of tens of milliseconds.) One can argue that it is preferable to have stimulus confounds than performance confounds, as the latter presumably affect brain dynamics in a more global and complex way. However, significant stimulus confounds are clearly also undesirable.

Indeed, if any method of generating a difference in consciousness across experimental conditions must be contaminated with some confounding factor or other, it would seem that there may be no single experimental design that could reveal the "pure," uncontaminated neural substrates of consciousness. However, a possible way forward is to *triangulate* on these substrates by combining the results from multiple experimental designs with disjoint sets of confounds into one overarching analysis, rather than counting on any one given design being the silver bullet. A simple illustration of the idea is as follows: if experimental design

A matches for stimulus properties but suffers from performance confounds, and design B matches for performance but suffers from stimulus confounds, then perhaps analysis of the combined data could reveal the common subset of neural activity that correlates with consciousness in both experiments. In the idealized case, this common subset of neural activity would be confound-free, since the confounds in design A that co-vary with consciousness are completely absent in design B, and vice versa. In other words, such an analysis approach could potentially reveal the "pure" neural basis of consciousness. In practice, the triangulation approach faces significant challenges, not the least of which is the possibility that the neural substrate of consciousness might interact with different confounding factors in distinct and non-linear ways, thus complicating the distillation of the "pure" substrate of consciousness across experiments. Nonetheless, we regard the general premise of the triangulation approach as promising and worthy of development in future work.

7. Conclusion

We have presented theoretical considerations for why it is crucial to control for task performance confounds in the neuroscientific study of consciousness. The feasibility and value of this approach is demonstrated by a growing body of literature in which performance-matched differences in awareness have been successfully isolated and computationally modeled. However, the performance-matching approach comes with a number of caveats and nuances that require careful consideration. A promising way forward may be to combine performance-matching approaches with other, complementary approaches so as to triangulate on the "pure," confound-free neural substrate of consciousness.

Acknowledgements

We thank for their comments on a previous version of this chapter Rachel Denison, J.D. Knotts, Enoch Lambert, Hakwan Lau, Anna Leshinskaya, Matthias Michel, Ian Phillips, Jason Samaha, the editors Felipe de Brigard and Walter Sinnott-Armstrong, and participants at the 2019 Summer Seminars in Neuroscience and Philosophy (SSNAP).

References

Aru, J., Bachmann, T., Singer, W., & Melloni, L. (2012). Distilling the neural correlates of consciousness. *Neuroscience and Biobehavioral Reviews*, *36*(2), 737–746.

Baars, B. J. (1997). In the theatre of consciousness. Global Workspace Theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, *4*(4), 292–309.

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. In Steven Laureys (Ed.), *Progress in Brain Research* (Vol. 150, pp. 45–53). Elsevier.

Baria, A. T., Maniscalco, B., & He, B. J. (2017). Initial-state-dependent, robust, transient neural dynamics encode conscious visual perception. *PLoS Computational Biology*, *13*(11), e1005806.

Bayne, T., Hohwy, J., & Owen, A. M. (2016). Are There Levels of Consciousness? *Trends in Cognitive Sciences*, *20*(6), 405–413.

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). https://doi.org/10.1523/ENEURO.0182-17.2017

Bickle, J. (2008). Real Reduction in Real Neuroscience: Metascience, Not Philosophy of Science (and Certainly Not Metaphysics!). In J. Hohwy & J. Kallestrup (Eds.), *Being Reduced* (pp. 34–51). Oxford University Press.

Block, N. (1995). On a confusion about a function of consciousness. *The Behavioral and Brain Sciences*, *18*(2), 227–247.

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences*, *9*(2), 46–52.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *The Behavioral and Brain Sciences*, *30*(5-6), 481–499; discussion 499–548.

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

Block, N. (2019). What Is Wrong with the No-Report Paradigm and How to Fix It. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2019.10.001

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the Higher-Order Approach to Consciousness. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2019.06.009

Brown, S., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(4), 587–599.

Chalmers, D. J. (2000). What is a neural correlate of consciousness? In T. Metzinger (Ed.), *Neural correlates of consciousness: Empirical and conceptual questions , (pp* (Vol. 350, pp. 17–39). Cambridge, MA, US: The MIT Press.

Cohen, M. A., Cavanagh, P., Chun, M. M., & Nakayama, K. (2012). The attentional requirements of consciousness. *Trends in Cognitive Sciences*, *16*(8), 411–417.

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8), 358–364.

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Penguin Publishing Group.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204–211.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492.

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: A Journal of Neurology*, *132*(Pt 9), 2531–2540.

Eriksen, C. W. (1960). Discrimination and learning without awareness: a methodological survey and evaluation. *Psychological Review*, *67*, 279–300.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *The American Psychologist*, *34*(10), 906–911.

Fleming, S. M. (2019). Awareness as inference in a higher-order state space. Retrieved from http://arxiv.org/abs/1906.00728

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1280–1286.

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

Giles, N., Lau, H., & Odegaard, B. (2016). What Type of Awareness Does Binocular Rivalry Assess? *Trends in Cognitive Sciences*, *20*(10), 719–720.

Godfrey-Smith, P. (2008). Reduction in Real Life. In J. Hohwy & J. Kallestrup (Eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford University Press.

Gorea, A., & Sagi, D. (2000). Failure to handle more than one internal representation in visual detection tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 12380–12384.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. 1966. *New York*.

Gross, S., & Flombaum, J. (2017). Does Perceptual Consciousness Overflow Cognitive Access? The Challenge from Probabilistic, Hierarchical Processes: Perceptual Consciousness and Cognitive Access. *Mind & Language*, *32*(3), 358–391.

Hannula, D. E., Simons, D. J., & Cohen, N. J. (2005). Imaging implicit perception: promise and pitfalls. *Nature Reviews. Neuroscience*, *6*(3), 247–255.

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *The Behavioral and Brain Sciences*, *9*(1), 1–23.

Irvine, E. (2012). *Consciousness as a Scientific Concept: A Philosophy of Science Perspective (Studies in Brain and Mind)* (2013 edition). Springer.

Kim, C.-Y., & Blake, R. (2005). Psychophysical magic: rendering the visible “invisible.” *Trends in Cognitive Sciences*, *9*(8), 381–388.

Knotts, J. D., Lau, H., & Peters, M. A. K. (2018). Continuous flash suppression and monocular pattern masking impact subjective awareness similarly. *Attention, Perception & Psychophysics*, *80*(8), 1974–1987.

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

Kouider, S., de Gardelle, V., Sackur, J., & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences*, *14*(7), 301–307.

Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1481), 857–875.

Kouider, S., Dehaene, S., Jobert, A., & Le Bihan, D. (2007). Cerebral bases of subliminal and supraliminal priming during reading. *Cerebral Cortex* , *17*(9), 2019–2029.

Kouider, S., Sackur, J., & Gardelle, V. de. (2012). [Review of *Do we still need phenomenal consciousness? Comment on Block*]. *Trends in cognitive sciences*, *16*(3), 140–141; author reply 141–142.

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

Lamme, V. A. F. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience*, *1*(3), 204–220.

Lamy, D., Salti, M., & Bar-Haim, Y. (2009). Neural correlates of subjective awareness and unconscious processing: an ERP study. *Journal of Cognitive Neuroscience*, *21*(7), 1435–1446.

Lau, H. C. (2008). Are we studying consciousness yet? In *Frontiers of Consciousness*. Oxford: Oxford University Press.

Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(49), 18763–18768.

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

Lloyd, D. A., Abrahamyan, A., & Harris, J. A. (2013). Brain-stimulation induced blindsight: unconscious vision or response bias? *PloS One*, *8*(12), e82828.

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

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 *The Cognitive Neuroscience of Metacognition* (pp. 25–66). Springer, Berlin, Heidelberg.

Maniscalco, B., & Lau, H. (2015). Manipulation of working memory contents selectively impairs metacognitive sensitivity in a concurrent visual discrimination task. *Neuroscience of Consciousness*, *2015*(1), niv002.

Maniscalco, B., & Lau, H. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness*, *2016*(1). https://doi.org/10.1093/nc/niw002

Maniscalco, B., McCurdy, L. Y., Odegaard, B., & Lau, H. (2017). Limited Cognitive Resources Explain a Trade-Off between Perceptual and Metacognitive Vigilance. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(5), 1213–1224.

Maniscalco, B., Peters, M. A. K., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception & Psychophysics*, *78*(3), 923–937.

Mathewson, K. E., Gratton, G., Fabiani, M., Beck, D. M., & Ro, T. (2009). To see or not to see: prestimulus alpha phase predicts visual awareness. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(9), 2725–2732.

Merikle, P. M., Smilek, D., & Eastwood, J. D. (2001). Perception without awareness: perspectives from cognitive psychology. *Cognition*, *79*(1-2), 115–134.

Michel, M. (2019). The Mismeasure of Consciousness: A problem of coordination for the Perceptual Awareness Scale. *Philosophy of Science*. https://doi.org/10.1086/705509

Michel, M., & Morales, J. (2019). Minority reports: Consciousness and the prefrontal cortex. *Mind & Language*, *30*, 1473.

Morales, J., Chiang, J., & Lau, H. (2015). Controlling for performance capacity confounds in neuroimaging studies of conscious awareness. *Neuroscience of Consciousness*, *2015*(1). https://doi.org/10.1093/nc/niv008

Morales, J., & Lau, H. (forthcoming). The Neural Correlates of Consciousness. In U. Kriegel (Ed.), *The Oxford Handbook of the Philosophy of Consciousness*. Oxford University Press.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450.

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.

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

Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *eLife*, *4*, e09651.

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

Phillips, I. B. (2011). Perception and Iconic Memory: What Sperling Doesn't Show. *Mind & Language*, *26*(4), 381–411.

Proust, J. (2013). *The Philosophy of Metacognition: Mental Agency and Self-Awareness*. OUP Oxford.

Rahnev, D. A., Bahdo, L., de Lange, F. P., & Lau, H. (2012). Prestimulus hemodynamic activity in dorsal attention network is negatively associated with decision confidence in visual perception. *Journal of Neurophysiology*, *108*(5), 1529–1536.

Rahnev, D. A., Maniscalco, B., Luber, B., Lau, H., & Lisanby, S. H. (2012). Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *Journal of Neurophysiology*, *107*(6), 1556–1563.

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., de Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, *14*(12), 1513–1515.

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

Rausch, M., & Zehetleitner, M. (2016). Visibility Is Not Equivalent to Confidence in a Low Contrast Orientation Discrimination Task. *Frontiers in Psychology*, *7*, 591.

Robinson, Z., Maley, C. J., & Piccinini, G. (2015). Is Consciousness a Spandrel? *Journal of the American Philosophical Association*, *1*(2), 365–383.

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

Rosenthal, D. M. (1993). State consciousness and transitive consciousness. *Consciousness and Cognition: An International Journal*, *2*(4), 355–363.

Rosenthal, D. M. (2008). Consciousness and its function. *Neuropsychologia*, *46*(3), 829–840.

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.

Ruby, E., Maniscalco, B., & Peters, M. A. K. (2018). On a "failed" attempt to manipulate visual metacognition with transcranial magnetic stimulation to prefrontal cortex. *Consciousness and Cognition*, *62*, 34–41.

Samaha, J., Barrett, J. J., Sheldon, A. D., LaRocque, J. J., & Postle, B. R. (2016). Dissociating Perceptual Confidence from Discrimination Accuracy Reveals No Influence of Metacognitive Awareness on Working Memory. *Frontiers in Psychology*, *7*, 851.

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., Switzky, M., & Postle, B. R. (2019). Confidence boosts serial dependence in orientation estimation. *Journal of Vision*, *19*(4), 25.

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.

Schwitzgebel, E. (2011). *Perplexities of Consciousness*. The MIT Press.

Sergent, C., & Dehaene, S. (2004). Neural processes underlying conscious perception: experimental findings and a global neuronal workspace framework. *Journal of Physiology, Paris*, *98*(4-6), 374–384.

Shaver, E., Maniscalco, B., & Lau, H. (2008). Awareness as Confidence. *Anthropology and Philosophy*, *9*(1/2), 58–65.

Shoemaker, S. (1981). Some Varieties of Functionalism. *Philosophical Topics*, *12*(1), 93–119.

Snodgrass, M., Bernat, E., & Shevrin, H. (2004). [Review of *Unconscious perception: a model-based approach to method and evidence*]. *Perception & psychophysics*, *66*(5), 846–867.

Spener, M. (n.d.). Consciousness, Introspection, and Subjective Measures. In U. Kriegel (Ed.), *The Oxford Handbook of the Philosophy of Consciousness*. Oxford University Press.

Taylor, E. (2016). Explanation and the Explanatory Gap. *Acta Analytica*, *31*(1), 77–88.

Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. F. (2015). No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. *Trends in Cognitive Sciences*, *19*(12), 757–770.

Weiskrantz, L. (1986). *Blindsight: A Case Study and Implications*. Oxford University Press.

Weiskrantz, L., Barbur, J. L., & Sahraie, A. (1995). Parameters affecting conscious versus unconscious visual discrimination with damage to the visual cortex (V1). *Proceedings of the National Academy of Sciences of the United States of America*, *92*(13), 6122–6126.