Confidence Tracks Consciousness*

Jorge Morales[1, 2] & Hakwan Lau[3]

[1] Department of Philosophy, Northeastern University
[2] Department of Psychology, Northeastern University
[3] RIKEN Center for Brain Science

5.1 Introduction

There is an obvious connection between feelings of confidence and the subjective experience of consciously perceiving something. When consciously perceiving something, one typically is at least somewhat confident about what that perceptual experience is about. Alternatively, when one doesn't have a conscious experience of something—even if one correctly perceives it unconsciously—one typically does not have a sense of confidence about what the perception is about. This link has often been used to justify the use of confidence ratings as an indirect measure of subjects' conscious awareness in perceptual tasks.[1]

Recently, Rosenthal (2019) discusses in depth these potential connections between confidence and consciousness but he ultimately rejects confidence as a useful, and in some cases even valid, measure of consciousness. Instead, he argues there are better alternatives to get at conscious experiences, such as direct subjective reports of awareness (i.e. subjects' sincere reports of perceiving something or of the degree of visibility of a stimulus).[2] Rosenthal concludes that "there can be little to favor confidence over subjective report as a measure of consciousness" (Rosenthal 2019, 264).

We agree with much of Rosenthal's analysis and we too share some of his concerns. In fact, we have also urged researchers to not equate confidence or metacognition with subjective experience (Fleming and Lau 2014; Maniscalco and Lau 2012; Morales, Odegaard, and Maniscalco 2019). Nevertheless, confidence may in fact offer a valuable window into consciousness. Metacognitive measures such as confidence ratings may offer important advantages over subjective ratings.

* We thank Cody Cushing, J. D. Knotts, Matthias Michel, Megan Peters, and Taylor Webb for valuable comments on an earlier manuscript.

[1] Most of our analysis is limited to visual experiences.

[2] In the literature, the terms "subjective reports" or "subjective ratings" sometimes include confidence ratings. In this chapter, confidence ratings should be understood as distinct from subjective reports.

5.2 Advantages in Using Confidence Ratings

To study consciousness in the laboratory, subjects typically perform a primary perceptual task (type I) and then provide a subjective judgment (type II) on that primary task. For instance, if asked to detect whether a stimulus was briefly presented on a screen or not, subjects' primary task is to respond "present" or "absent" (normally by pressing a key). Then, subjects may be asked to provide a *subjective report* about the visibility of the stimulus during the type I task. For example, they may be asked to respond "seen" or "guess". The answers may also be more fine-grained, as in the Perceptual Awareness Scale (PAS) (Ramsøy and Overgaard 2004), which introduces four levels or degrees of awareness: (0) No experience, (1) Brief glimpse, (2) Almost clear experience, and (3) Clear experience. These kinds of subjective reports are commonly used in consciousness research and they are meant to be direct reports of subjects' experience in each experimental trial. Alternatively, instead of providing a subjective report, subjects may make a *confidence judgment* in the correctness of their type I response. This is typically done by pressing a key that maps onto some scale that tracks different levels of confidence (e.g., low vs high confidence, or a 4-point scale that goes from "no confidence" to "certain", etc.).

The central advantage of using confidence ratings over subjective ratings of visibility is their ease of clear definition and instruction. Unlike subjective reports, confidence can be clearly defined as the subjective probability of being correct in the primary task (Norman and Price, 2015). This definition allows experimenters to treat *subjective* confidence ratings as a somewhat *objective* measure of metacognitive sensitivity. That is, subjective reports of visibility are about subjective experiences, which are inaccessible to the experimenter. However, when using confidence ratings, there is a truth of the matter: subjects' confidence ratings either predict or not the correctness of their type I responses, thus objectively tying confidence ratings to task performance. This relationship can be used to estimate subjects' *metacognitive sensitivity* (Fleming and Lau 2014). For instance, an ideal observer would rate accurate responses in the primary task with high confidence and inaccurate responses with low confidence. These confidence ratings are objectively correct. In contrast, a low confidence rating after a correct response or a high confidence rating after an incorrect response are objectively incorrect. Subjects' metacognitive behavior can be compared against this ideal to measure their metacognitive sensitivity (Fleming 2017; Galvin et al. 2003; Maniscalco and Lau 2016).

Relatedly, an advantage of the objectivity of confidence ratings is that subjects can receive feedback on their metacognitive performance, and thus can be trained to improve the accuracy of their confidence ratings (Carpenter et al. 2019). Animals can also be trained to rate confidence, even though verbal instruction is not possible (Kiani and Shadlen 2009; Kornell, Son, and Terrace 2007; Smith, Shields, and Washburn 2003; Smith, Couchman, and Beran 2014)(Kiani & Shadlen 2009; Kornell *et al.* 2007; Smith *et al.* 2003, 2014). All this is relatively difficult if not impossible to achieve with subjective reports of conscious awareness.

Perhaps the most important benefit provided by the clear definition and instruction of confidence ratings is that they can remove a significant amount of measurement noise. When using subjective reports, participants may interpret quite differently what "seeing something" or "not seeing anything" means. This reflects the so-called *criterion content problem* (Kahneman 1968). Subjective reports about a stimulus necessarily are the result of certain unspecified criteria of what to focus on when reporting back the experience. Put simply, what counts as "seeing something" or "not seeing anything" is often not clear to subjects, and the criterion of what counts as such may change across subjects or across trials for the same subject.

This problem arises even when using *prima facie* clearly defined scales for subjective reports of visibility, such as the PAS (Ramsøy and Overgaard 2004). Even though it provides labels for each of the levels of the scale, the PAS is not without problems (Michel 2019). For example, the PAS cannot provide any guidance as to what criterion to use for choosing each level. Does a "brief glimpse" mean being aware of *anything at all* or of *some meaningful feature* of the stimulus? Having a vague experience of something being present on the screen versus having a vague experience of a left-tilted grating being on the screen are quite different, and yet the scale itself does not constrain a consistent usage. (This is true even if experimenters try to specify what each scale is supposed to mean.) One subject might interpret a vague experience of an unspecified content as "no experience" and another might interpret it as a "brief glimpse". In most experiments that probe conscious awareness, stimuli are at threshold or somehow degraded (e.g., low contrast, fast presentation, masking, distracted attention, etc.). Under these conditions, subjects may still see *something*. The option "no awareness" may be too strong and it may be interpreted very differently by different subjects, in different tasks. In an odd sense, one always 'sees' something, even with one's eyes closed ("seeing darkness"?). This sense of understanding seeing may be odd, but some subjects may well hear it that way. When we study a large group of people these problems do occur, and the use of subjective reports leaves open the door for these odd or inconsistent behaviors.

In stark contrast, confidence can be expressed on a well-calibrated and meaningful scale.[3] Probability in terms of percentage likelihood is comparable and equally applicable across different tasks (e.g., detection, discrimination, recognition, etc.). Defined this way (instead of on an arbitrary scale, e.g., high vs low), confidence judgments can reflect the subjective probability of one's being correct in a task, no matter the nature of the task. That is, the questions "how confident are you?" or "how likely are you to be correct?" remain the same for different types of tasks. This has allowed for comparisons of type II behavior between perceptual domains that have very different phenomenologies (Rouault et al. 2018), such as different perceptual modalities (de Gardelle, Le Corre, and Mamassian 2016; Faivre et al. 2017), or between domains such as visual perception and memory (Fitzgerald, Arvaneh, and Dockree 2017; McCurdy et al. 2013; Morales, Lau, and Fleming 2018). Even transferring metacognitive sensitivity training from one domain to another is possible (Carpenter et al. 2019). In contrast, subjective ratings are most applicable to single stimulus detection (and perhaps discrimination), and the response options obviously need to be modified depending on the nature of the task and the stimuli involved (e.g. comparing two sets of different stimuli, detecting if something has changed or is missing, etc.).

Last but not least, one often neglected consideration is socio-strategic. The study of confidence and metacognition is a burgeoning field within mainstream cognitive psychology and cognitive neuroscience. There is rigorous work on both computational (Fleming and Daw 2017; van den Berg et al. 2016) and psychophysical modeling (Fleming 2017; Maniscalco and Lau 2016), as well as neuronal electrophysiology (Kiani and Shadlen 2009; Miyamoto et al. 2017; Stolyarova et al. 2019; Miyamoto et al. 2018). There is hardly any such equivalence for the study of 'subjective visibility'. This has little to do with substantive theoretical considerations, but it is no less important. To the extent that the two kinds of measures are similar in most cases, as we will argue in section 4, such consideration becomes relevant. Science is very much a social activity. The standards of rigor, funding and job availability matter, and they largely depend on

[3] This, of course, does not mean that subjects cannot be biased (e.g., under- or overconfident) in how they use confidence ratings. However, the meaning of each level (i.e., the probability of being correct) is well-defined, even if thinking about subjective probabilities is hard or if mapping subjective probabilities onto a specific scale could create some noise.

which peer groups one belongs to. And this matters not just in terms of personal benefit but also for the longevity and development of the field. Making the scientific study of consciousness relevant to the study of confidence and metacognition is strategically appealing.

5.3 Problems with Confidence Ratings

The advantages discussed in the previous section are, however, not decisive by themselves. They have to be weighed up against other factors, such as potential caveats about the use of confidence ratings.

One such caveat raised by Rosenthal (2019) concerns the possibility of enjoying conscious experiences without confidence. This is the case in peripheral vision. One may not be confident of *what* one is seeing in the periphery of the field of vision, and yet enjoy a distinct conscious experience of seeing *something*. We basically agree with this description of the phenomenology. However, it is important to distinguish between detection (Is there something?) and discrimination (What is it? Is it A or B?). Our discriminative ability is relatively poor outside of the focus of attention (Braun et al. 1999) (Braun *et al.* 1999). Accordingly, it is not surprising to have low confidence judgments of discrimination in the periphery. In fact, optimal metacognizers are expected to rate their discriminations in the periphery with low confidence. In contrast, detection in the periphery is known to be liberal (i.e. subjects tend to report often that they detected something) (M. K. Li, Lau, and Odegaard 2018; Odegaard et al. 2018; Solovey, Graney, and Lau 2015). Thus, when confidence concerns detection rather than discrimination, subjects are *more* likely to be confident after seeing something in the periphery. On these trials, they are likely to judge that they are fairly sure that they see something (confidence for detection), even if they are unsure of what they see (confidence for discrimination). So, Rosenthal's suggestion that there is a dissociation between confidence and subjective experience in the periphery should be limited to some kinds of tasks only. For confidence ratings during detection tasks, confidence and awareness seem to go hand in hand.

Another possible dissociation between consciousness and confidence is when confidence is not based on a subjective perceptual experience. Rosenthal discusses the somewhat complicated case of type II blindsight. Blindsight is a condition in which patients with a lesion in the visual cortex deny being consciously aware of stimuli presented in a specific region of their visual field. In type II blindsight, blindsight patients claim to "feel" some change within their blindfield (e.g., movement) but they also insist they know this not because of having a normal visual experience (i.e., they just "feel" it) (Brogaard 2014; Foley and Kentridge 2015; Foley 2015; Macpherson 2015). Rosenthal takes blindsight patients' denial of having a normal visual experience as evidence that their "nonvisual awareness is not perceptual in any way; it is best seen as a type of cognition" (2019, 262). While it seems clear that blindsight patient's experience is not normal (e.g., it doesn't feel the same way as their normal visual field), this need not entail it is not visual (e.g., presumably, if they closed their eyes the "feeling" would go away). Even conceding type II blindsight does not entail that patients enjoy *visual* experiences, it is not clear that we should thereby infer they are not reporting a conscious experience. And if they do and they base their confidence on this subjective conscious experience (even if it is not visual), the connection between confidence and consciousness might still survive—or at least it would not be imperiled to the degree and for the reasons suggested by Rosenthal.

Type II blindsight, however, is a complex case, ultimately difficult to analyze. Perhaps a much simpler scenario can highlight Rosenthal's worries about confidence being disconnected from conscious experiences without the vicissitudes of type II blindsight. 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 second form of dissociation could become a real problem, and is in part why we too recommend researchers not to equate consciousness with confidence (Fleming and Lau 2014; Maniscalco and Lau 2012; Morales, Odegaard, and Maniscalco 2019). However, the fact that this kind of dissociation can take place does not entail that confidence is not, in general, a good indicator of consciousness. The degree to which confidence reflects subjective experience ultimately depends on the degree to which confidence is exhaustively driven by conscious perceptual information. Thus, to the extent that we can rule out non-perceptual sources of confidence, confidence does track subjective experience very closely.

5.4 Similarity between Confidence and Subjective Ratings

Despite the advantages of using confidence over subjective reports of awareness (section 2) and the aforementioned dissociations between the two (section 3), in practice they produce very similar experimental results. These important, yet theoretical differences, do not appear to be significant enough to produce behavioral differences in the laboratory—at least not with our current methods.

Behaviorally, visibility and confidence ratings seem to produce similar results. For example, Peters & Lau (Peters and Lau 2015) obtained qualitatively identical results in a masking experiment regardless of whether they asked subjects to rate their confidence or to judge the visibility of the stimuli. Even researchers who have found (rather small) differences between subjective reports and confidence ratings in conditions of very low contrast (e.g., Rausch and Zehetleitner 2016; Zehetleitner and Rausch 2013) admit that "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" (Zehetleitner and Rausch 2013, 1423).

But a stronger point can be made about the close connection between confidence ratings and subjective reports. The underlying neural dynamics and neural mechanisms supporting different types of subjective reflection on one's experience largely overlap. First, certain features of the brain's global dynamics affect visibility and confidence ratings in a similar way. Spontaneous low frequency brain oscillations (<30 Hz) affect (or perhaps reflect) neuronal excitability and, with it, performance and type II ratings during psychophysical tasks (Samaha et al. 2020). In particular, two recent studies found that low frequency oscillations with lower prestimulus power (i.e., oscillations of lower magnitude right before the presentation of a stimulus) biased observers to report both higher confidence and higher subjective visibility (Benwell et al. 2017; Samaha, Iemi, and Postle 2017). Second, despite stark task differences and radically different ways of probing consciousness, many studies using different types of neuroimaging techniques across different species have consistently found astonishingly similar neural correlates of consciousness. Prefrontal cortex (PFC) (often very specific areas in dorsolateral and orbitofrontal PFC) has been found to support subjective reports of awareness (Del Cul et al. 2009; Lau and Passingham 2006), visibility ratings (Rounis et al. 2010) and confidence ratings alike [in both animals (Mendoza-

Halliday and Martinez-Trujillo 2017) and humans (Cortese et al. 2016; Fleming, Huijgen, and Dolan 2012; Morales, Lau, and Fleming 2018)]. Importantly, these findings likely reflect the underlying perceptual experience rather than the mere act of reporting it (Michel and Morales 2020).

Despite these widespread similarities, Rosenthal cites research showing that subjective reports and confidence ratings have different neural activity profiles. In particular, he appeals to a study by Li et al. (2014) to argue that we have reasons "to expect that confidence ratings and subjective awareness likely reflect different psychological processes, at least to some extent" (Rosenthal 2019, 259). This, however, should not matter for using confidence as a *proxy* for consciousness. Even if they are different psychological processes *to some extent*—as we admit they are—we can use one to learn about the other (see section 5). We dispute, however, that Li at colleagues' results support defending a *significant* difference between the neural profiles of subjective reports and confidence ratings or, more importantly, a difference that is significant for the study of consciousness. Our reasons are somewhat technical, but we think they are worth reviewing because of their ultimate importance for the neuroscientific study of consciousness in general.

In Li et al.'s study, subjects saw Gabor patches oriented to the left or to the right in each trial. They had to answer three sequential questions: (1) Was the Gabor patch pointing left or right? (2) Did you see the stimulus or not? (3) How confident are you about your answer to question (2)? Subjects' magnetoencephalographic (MEG) activity was recorded throughout the experiment. MEG activity correlated with subjective awareness (i.e., the answer to question 2) peaked between .5 and 1.5 seconds after stimulus offset and it covered widespread frontoparietal and temporal areas. In contrast, MEG activity correlated with confidence ratings (i.e., the answer to question 3) peaked at .5 seconds in frontoparietal areas and dissipated shortly after. Li and colleagues concluded that compared with subjective awareness, confidence is associated with relatively transient MEG activity.

First, we should point out that Li et al.'s results confirm the frontal localization shared by subjective reports and confidence ratings we discussed above. However, it is important to note that subjects were asked to answer a non-standard confidence question. In most experiments using confidence ratings subjects are asked to rate their confidence in their type I decision. Li et al., in contrast, asked subjects to evaluate their confidence in their subjective report (a type II evaluation of a type II question). While this may be an interesting and valid approach, the peculiarity of the procedure makes comparisons with other studies hard to evaluate.

More problematically, the alleged difference in the MEG profile of subjective reports and confidence ratings is hard to evaluate because the underlying analysis suffers from an important confound: task performance is not matched between aware and unaware conditions. A crucial step when comparing aware vs unaware (or high vs low confidence) neural data is to ensure performance in the main task is matched; otherwise, instead of comparing the neural correlates of consciousness one risks just detecting differences in perceptual processing (Lau 2008; Morales, Odegaard, and Maniscalco 2019). While perhaps tempting, one cannot attempt to match performance "artificially" by simply analyzing the correct trials of the two conditions of interest (Morales, Chiang, and Lau 2015). It may appear as a tempting solution because one could think that correct aware and correct unaware trials have a matched performance (100% accuracy for both!). But, crucially, the perceptual signal that gives rise to correct aware trials is most certainly stronger than the perceptual signal that allows for a correct answer in an unaware trial. In the former, the internal perceptual response is more likely to be high; in the latter, however, the internal perceptual response is likely to be low. This entails that a larger proportion of unaware correct answers is the product of chance rather than perceptual discrimination (even when you don't see the stimulus, you have a 50/50 chance of guessing correctly the answer in the main task). Despite these known problems surrounding performance-matching corrections, the analyses in Li et al. (2014; their figure 2C) incorporate comparisons between aware correct vs unaware correct MEG activity. This kind of comparison that overlooks true performance (and in consequence internal response) obscures the underlying nature of aware and unaware neural activity. Thus, comparing these results to those pertaining to confidence becomes extremely difficult—and potentially invalid.

The overwhelming behavioral and neural similarity between subjective reports and confidence ratings should not be particularly surprising. After all, they are both subjective assessments and in the case of confidence ratings, they are likely to be driven to a large extent by conscious experiences themselves (see section 5). In fact, they are so similar that many of Rosenthal's criticisms against confidence ratings apply almost equally to subjective reports of awareness as well. For example, consider his argument against using confidence ratings because they offer no benefit over subjective reports in cases of complete lack of confidence. "For confidence to be a useful indicator of consciousness, subjects would have to distinguish total lack of confidence from very slight confidence," Rosenthal thinks.

"It is unlikely that subjects would be more accurate in drawing that distinction than in distinguishing minimal awareness from complete absence of awareness." (Rosenthal 2019, 258) We agree that subjects are asked to make this distinction; especially in detection experiments where subjects have to evaluate their confidence in whether they saw something at all or not. But at least some confidence rating scales have a "guess" option at the lower end (Dienes et al. 1995; Dienes and Seth 2010; Wierzchoń, Asanowicz, and Paulewicz 2012).[4] Forcing subjects to distinguish a complete absence of confidence (i.e., full guessing) from a minimal degree of confidence should be possible with these scales, satisfying Rosenthal's demand. And even though scales with "guess" or "no confidence" options have been deemed problematic (Norman and Price, 2015), very similar criticisms about the interpretability of this option have been raised against subjective reports too (Michel 2019). Moreover, even if Rosenthal were right that using confidence ratings cannot make subjects *more* accurate (but see the arguments from section 2), it is not clear that using confidence ratings would make subjects *less* accurate. As he admits, "subjective reports can be biased [...] and may not always reflect subjective awareness with total accuracy" (Rosenthal 2019, 258). In fact, the presence of response biases in subjective reports of awareness may be problematic (Phillips 2016) and may even be unavoidable (Peters, Ro, and Lau 2016).

In the end, we think that the advantages of using confidence over subjective reports, in addition to their strong similarities and very small and subtle differences, should favor the use of confidence ratings. But we do not advocate this as a strict dogma. If a specific situation indicates confidence may be problematic (e.g., prior knowledge may affect the results, the subject population have self-esteem traits that might unduly inflate or deflate confidence ratings, etc.), then we do not discourage the use of subjective reports (e.g. PAS or some other scale). After all, many of the benefits and problems of confidence and subjective reports are similar. The decision to use one or another scale is mostly methodological: depending on the specific design and goal of a study, using confidence ratings might not be the best course of action for that particular case. However, this does not make confidence less desirable as a tool for studying consciousness and, as we've argued above, confidence is preferable over subjective reports in a vast number of cases.

[4] Even though these scales have been used in assessing awareness of artificial grammars, nothing prevents us from using them in visual tasks.

5.5 Quality Space Theory and Confidence

We now turn to a more conceptual question: what is the link between consciousness and confidence? In other words, why does confidence seem to reflect consciousness, at least most of the time?

Philosophers and scientists alike often claim that there are associations between consciousness and cognitive functions, e.g., one needs to be consciously aware of certain information to exercise cognitive control over it, initiate voluntary action, exercise rational thought, and display flexible behavior (Dehaene et al. 2014; Tye 1996). These claims, however, are also sometimes challenged on both empirical (Koizumi, Maniscalco, and Lau 2015; Lau and Passingham 2007; van Gaal et al. 2008; van Gaal, de Lange, and Cohen 2012) and philosophical (Robinson, Maley, and Piccinini 2015) grounds. More generally, there is a serious technical challenge experimenters face when studying the functions of consciousness. As noted in the previous section, a problem that is hard to overcome is that perceptual signals are often confounded with consciousness. In the typical case, strong perceptual signals are correlated with conscious perception and weak perceptual signals are correlated with unconscious perception. With stronger perceptual signals more cognitive functions are trivially expected to be exercised (Block 2019; Phillips and Morales 2020). Thus, without properly matching perceptual signals (e.g. by matching task performance), simply looking at the functions of consciousness that are lost during unconscious perception may tell us little about consciousness per se (Morales, Odegaard, and Maniscalco 2019).

It may be tempting to think that consciousness and confidence (metacognition in particular) may be similarly confounded. Recall that here, by metacognition we understand one's ability to rate confidence meaningfully (i.e. to rate confidence in a way that closely tracks one's performance in a given task). Is it possible that consciousness's link to metacognition is just as tenuous as its link to other higher cognitive functions? Rosenthal thinks this is the case. According to him, although confidence has considerable utility (e.g. it informs rational decision making), a psychological state's being conscious does not add any utility to the state (Rosenthal 2012; Rosenthal 2008). Therefore, consciousness and confidence cannot be linked in any strong sense (Rosenthal 2019).

There may be, however, a more substantive link between consciousness and metacognition than Rosenthal allows. In fact, a variation of Rosenthal's own higher-order thought theory and mental quality space theory might provide important clues into this link. In a nutshell, we will argue that without consciousness one should not expect to do metacognition nearly as well. In other words, consciousness does inform our confidence judgments in a significant way.

According to quality space theory, "mental qualities are properties of states in virtue of which an organism responds to a range of perceptible properties" (Rosenthal 2005, 202). Mental qualities are defined "by their position in a quality space that's homomorphic to the quality space of the perceptible properties accessible to that modality" (*idem*). This entails that an organism's quality space is entirely determined by the most fine-grained discriminations it can make. To find out the limit of an organism's discrimination ability, one can test experimentally for just noticeable differences (JNDs). For example, in the case of color, one would use color stimuli that are so close physically that they would be perceptually indistinguishable if they were any closer. Importantly, to discriminate these stimuli from one another, the organism must be able to be in psychological states that differ correspondingly. This is how a homomorphism between stimulus properties and mental qualities is obtained. The quality space "that represents the stimuli an individual can discriminate will also represent the similarities and differences among the perceptual states in virtue of which such discriminations are possible for that individual" (Rosenthal 2015, 38). Importantly, according to Rosenthal, the psychological states that make these discriminations possible need not be conscious.

Now, consider the following toy example in which we try to characterize the mental quality space of single numerical digits, namely, a quality space of the visual similarities and differences between Arabic numerals. By running multiple pairwise discriminations between the digits, we can work out subjects' digit discrimination ability in terms of JNDs. Thus, we can put subjects' digit mental qualities on a quality space, such that the pairwise distance between the digits reflects their discriminability in JND units. Accordingly, '3' and '5' may be relatively close because they are somewhat more easily confused with each other. The distance in the quality space between '3' and '5', then, will be smaller compared to the distance of either of them to '1', because it's harder to confuse them with '1'. In contrast, '1' and '7' will be close to each other, and more distant from '3' and '5', because it is harder to discriminate 1's from 7's. The mental quality of each percept is defined by its position on this quality space.

According to higher-order thought theory, when one sees a stimulus consciously—via a suitable higher-order thought that represents the first-order perceptual state that represents the stimulus—one also becomes aware of the stimulus's quality (Rosenthal 2005). In Rosenthal's view (personal communication), the subjects do not necessarily have an explicit grasp of the detailed mental quality space. In other words, higher-order thoughts do not need to explicitly represent the percept's precise position on the mental quality space as such.

Let us assume for a moment, however, that in virtue of being conscious of the qualities of our percepts we knew their relative positions on the mental quality space. This would clearly be a useful piece of knowledge to possess. Imagine we ask you to name a digit that was quickly presented on a screen. If we told you that your initial answer of, say, '5' is wrong, in your second try you may well be more likely to say '3' than '1'. But this will be because you know which stimuli are fewer JNDs away from your initial answer than others. Similarly, in a two-choice discrimination, if we ask you whether the digit was '5' or '3', you may say '5' with limited confidence. But if the question was whether the digit was '5' or '1', you may choose '5' with a much higher confidence. Importantly, you make these confidence judgments based on your grasp of the distance in quality space between the two candidate digits.

It is possible that a subject with no awareness of the positions of these qualities on the mental quality space could adopt a similar strategy based on a space of the *physical* similarity between the stimuli. But it is not clear if ordinary, untrained subjects would do that. Whereas in conscious perception one seems to just make these metacognitive judgements without any such explicit strategy of searching through a space. Notably, if a subject made these discriminations based on their awareness of their percepts' positions on the mental quality space rather than on a physical similarity space, we should expect their metacognitive performance to be superior. This is so because, as explained above, mental quality spaces are determined by one's very own perceptual abilities. In this sense, knowing the position of a percept on one's quality space is already a kind of self-knowledge that can be leveraged by metacognition.

Even though Rosenthal does not think we have an explicit grasp of the mental quality space in detail, in his view, higher-order thoughts may conceptualize the contents of first-order states in terms of the quality space positions in a relatively coarse-grained manner. For instance, "for colors broadly taxonomized, we all recognize that orange is closer, at least in respect of hue, to both red and yellow than it is to either green or blue. [...] These broad-stroked similarity relations allow one to construct a relatively coarse-grained space of colors [...] which capture these relations of similarity and difference" (Rosenthal 2015, 37). But this coarse-grained knowledge allows us to know, even if just roughly, the percept's relative position on the mental quality space. And this is consistent with the fact that metacognition usually isn't perfect: we might not know exactly or with infinite fineness of grain the percept's position on the mental quality space. However, without awareness, we should expect metacognition to be even worse (barring the kind of non-trivial, indirect strategy using a physical similar space mentioned above).

This kind of metacognitive benefit from consciousness can be confirmed in blindsight patients. Persaud and colleagues (Persaud et al. 2011) tested blindsight patient GY's metacognitive ability in both his blind and his normal hemifields. Stimuli were titrated to ensure that performance was matched in both his normal and blind hemifields. After providing a first-order response about the position on the screen of a target stimulus, GY could choose to get paid either via a coin flip (i.e. he had a 50/50 chance to earn/lose 50 cents regardless of his performance) or he could choose to be paid based on the correctness of his response (he would earn 50 cents if his response was correct and he would lose the same amount if it was not). This "no loss" post-decision wagering system (Dienes and Seth 2010; Persaud, McLeod, and Cowey 2007) essentially tracked GY's confidence in his own response. Although GY was not metacognitively "blind" in his blind hemifield (i.e. his wagers tracked to some extent his correct/incorrect responses), it was far inferior than his metacognitive sensitivity in his normal hemifield. In other words, consciousness seems to come with some added utility: it improves metacognition.

This case suggests that higher-order representations may actually code positional information with respect to the mental quality space in a much more fine-grained way than Rosenthal seems to allow. This would provide an account of our knowing what it is *like* to see a number '5' when we consciously see it: it is a little bit *like* a '3', but very much *unlike* a '1', etc. This similarity profile with respect to all other *possible* percepts within a quality space reflects the fine-grained richness of subjective perception. If this is correct, conscious seeing might constitutively involve grasping these similarity relations and, in turn, being available for metacognition.

These points do not establish that consciousness is necessary or sufficient for metacognition. One may know the position of a stimulus on the mental quality space and yet fail to make use of such information. Or one may use other strategies to make metacognitive confidence judgments. But here we suggest that there is a close link between consciousness and metacognitive mechanisms. This may explain why higher-order awareness tends to lead to superior metacognitive performance in a non-trivial way. A consciousness advantage for metacognitive sensitivity emerges not just because conscious signals tend to be stronger. Rather, blindsight and the digit quality space example point towards the existence of an inherent mechanistic advantage for metacognition when one perceives a stimulus consciously.

5.6 Concluding Remarks

We argued that the use of confidence in assessing consciousness is not theoretically arbitrary. While consciousness and confidence are definitely not identical, there are good reasons to think they are closely linked. This allows researchers interested in studying consciousness to use confidence ratings as reliable proxies of subjective ratings of consciousness. We discussed some problems with the use of confidence ratings, but many of these also apply to subjective ratings. One exception may be the case of confidence being informed by non-perceptual or prior knowledge. When such possibility cannot be ruled out, subjective ratings may be a good alternative. But in most other cases, the advantages of confidence ratings we outlined here outweigh their limitations.

We are confident that these will not be the last words on the matter. Methodological questions of the kind we discussed here can be expected to be solved only in the very long run. But this is exactly why this kind of friendly, multidisciplinary debate is so valuable. We are immensely grateful to David Rosenthal for capturing the relevant problems and for stimulating our thoughts, like he has done on practically every other issue related to consciousness.

References

Benwell, Christopher S Y, Chiara F Tagliabue, Domenica Veniero, Roberto Cecere, Silvia Savazzi, and Gregor Thut. 2017. "Prestimulus EEG Power Predicts Conscious Awareness but Not Objective Visual Performance." *eNeuro* 4 (6: ENEURO.0182–17.2017. doi:10.1523/ENEURO.0182-17.2017.

Block, Ned. 2019. "What Is Wrong with the No-Report Paradigm and How to Fix It." 23 (12): 1003–13. doi:10.1016/j.tics.2019.10.001.

Braun, J, D K Lee, Laurent Itti, and Christof Koch. 1999. "Attention Activates Winner-Take-All Competition Among Visual Filters." *Nature Neuroscience* 2 (4): 375–81. doi:10.1038/7286.

Brogaard, Berit. 2014. "Consciousness and Cognition." *Consciousness and Cognition*, October. 1–12. doi:10.1016/j.concog.2014.09.017.

Carpenter, Jason, Maxine T Sherman, Rogier A Kievit, Anil K Seth, Hakwan Lau, and Stephen M Fleming. 2019. "Domain-General Enhancements of Metacognitive Ability Through Adaptive Training." *Journal of Experimental Psychology: General* 148 (1): 51–64. doi:10.1037/xge0000505.

Cortese, Aurelio, Kaoru Amano, Ai Koizumi, Mitsuo Kawato, and Hakwan Lau. 2016. "Multivoxel Neurofeedback Selectively Modulates Confidence Without Changing Perceptual Performance." *Nature Communications* 7: 13669. doi:10.1038/ncomms13669.

de Gardelle, Vincent, François Le Corre, and Pascal Mamassian. 2016. "Confidence as a Common Currency Between Vision and Audition." *PLoS ONE* 11 (1): e0147901–11. doi:10.1371/journal.pone.0147901.

Dehaene, Stanislas, Lucie Charles, Jean-Remi King, and Sébastien Marti. 2014. "Toward a Computational Theory of Conscious Processing." *Current Opinion in Neurobiology* 25 (April): 76–84. doi:10.1016/j.conb.2013.12.005.

Del Cul, A, Stanislas Dehaene, P Reyes, E Bravo, and A Slachevsky. 2009. "Causal Role of Prefrontal Cortex in the Threshold for Access to Consciousness." *Brain* 132 (9): 2531–40. doi:10.1093/brain/awp111.

Dienes, Zoltán, and Anil Seth. 2010. "Gambling on the Unconscious: a Comparison of Wagering and Confidence Ratings as Measures of Awareness in an Artificial Grammar Task." *Consciousness and Cognition* 19 (2): 674–81. doi:10.1016/j.concog.2009.09.009.

Dienes, Zoltán, Gerry Altmann, Liam Kwan, and Alastair Goode. 1995. "Unconscious Knowledge of Artificial Grammars Is Applied Strategically." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21 (5): 1322–38.

Faivre, Nathan, Elisa Filevich, Guillermo Solovey, Simone Kuhn, and Olaf Blanke. 2017. "Behavioural, Modeling, and Electrophysiological Evidence for Supramodality in Human Metacognition." *The Journal of Neuroscience* 38 (2): 263–77. doi:10.1523/JNEUROSCI.0322-17.2017.

Fitzgerald, Lisa M, Mahnaz Arvaneh, and Paul M Dockree. 2017. "Consciousness and Cognition." *Consciousness and Cognition* 49 (March): 264–77. doi:10.1016/j.concog.2017.01.011.

Fleming, Stephen M. 2017. “Hierarchical Bayesian Estimation of Metacognitive Efficiency from Confidence Ratings.” *Neuroscience of Consciousness* 3 (1): 1–14. doi:10.1093/nc/nix007.

Fleming, Stephen M, and Hakwan Lau. 2014. “How to Measure Metacognition.” *Frontiers in Human Neuroscience* 8 (July): 443. doi:10.3389/fnhum.2014.00443.

Fleming, Stephen M, and Nathaniel D Daw. 2017. “Self-Evaluation of Decision-Making: a General Bayesian Framework for Metacognitive Computation.” *Psychological Review* 124 (1): 91–114. doi:10.1037/rev0000045.

Fleming, Stephen M, Josefien Huijgen, and Raymond J Dolan. 2012. “Prefrontal Contributions to Metacognition in Perceptual Decision Making.” *The Journal of Neuroscience* 32 (18): 6117–25. doi:10.1523/JNEUROSCI.6489-11.2012.

Foley, Robert. 2015. “Consciousness and Cognition.” *Consciousness and Cognition* 32: 56–67. doi:10.1016/j.concog.2014.09.005.

Foley, Robert, and Robert W Kentridge. 2015. “Consciousness and Cognition.” *Consciousness and Cognition* 32: 1–5. doi:10.1016/j.concog.2015.01.008.

Galvin, Susan J, John V Podd, Vit Drga, and John Whitmore. 2003. “Type 2 Tasks in the Theory of Signal Detectability: Discrimination Between Correct and Incorrect Decisions.” *Psychonomic Bulletin & Review* 10 (4): 843–76.

Kahneman, D. 1968. “Method, Findings, and Theory in Studies of Visual Masking.” *Psychological Bulletin* 70 (6): 404–25. doi:10.1037/h0026731.

Kiani, R, and M N Shadlen. 2009. “Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex.” *Science* 324 (5928): 759–64. doi:10.1126/science.1169405.

Koizumi, Ai, Brian Maniscalco, and Hakwan Lau. 2015. “Does Perceptual Confidence Facilitate Cognitive Control?” *Attention, Perception, & Psychophysics* 77 (4): 1295–1306. doi:10.3758/s13414-015-0843-3.

Kornell, N, L K Son, and H S Terrace. 2007. “Transfer of Metacognitive Skills and Hint Seeking in Monkeys.” *Psychological Science* 18 (1): 64–71. doi:10.1111/j.1467-9280.2007.01850.x.

Lau, H. 2022. *In Consciousness We Trust.* Oxford University Press.

Lau, Hakwan. 2008. “Are We Studying Consciousness Yet?” In *Frontiers of Consciousness*, edited by Lawrence Weiskrantz and Martin Davies, 245–58. Oxford University Press. doi:10.1093/acprof:oso/9780199233151.003.0008.

Lau, Hakwan, and R E Passingham. 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): 18763–68. doi:10.1073/pnas.0607716103.

Lau, Hakwan, and R E Passingham. 2007. “Unconscious Activation of the Cognitive Control System in the Human Prefrontal Cortex.” *The Journal of Neuroscience* 27 (21): 5805–11. doi:10.1523/JNEUROSCI.4335-06.2007.

Li, Musen Kingsley, Hakwan Lau, and Brian Odegaard. 2018. “An Investigation of Detection Biases in the Unattended Periphery During Simulated Driving” *Attention, Perception, & Psychophysics*, 80: 1325-1332. doi:10.3758/s13414-018-1554-3.

Li, Q, Z Hill, and B J He. 2014. "Spatiotemporal Dissociation of Brain Activity Underlying Subjective Awareness, Objective Performance and Confidence." *The Journal of Neuroscience* 34 (12): 4382–95. doi:10.1523/JNEUROSCI.1820-13.2014.

Macpherson, Fiona. 2015. "The Structure of Experience, the Nature of the Visual, and Type 2 Blindsight." *Consciousness and Cognition* 32 (March): 104–28. doi:10.1016/j.concog.2014.10.011.

Maniscalco, Brian, and Hakwan Lau. 2012. "A Signal Detection Theoretic Approach for Estimating Metacognitive Sensitivity from Confidence Ratings." *Consciousness and Cognition* 21: 422–30.

Maniscalco, Brian, and Hakwan Lau. 2016. "The Signal Processing Architecture Underlying Subjective Reports of Sensory Awareness." *Neuroscience of Consciousness* 2016 (1): 292. doi:10.1093/nc/niw002.

McCurdy, Li Yan, Brian Maniscalco, Janet Metcalfe, Ka Yuet Liu, Floris P de Lange, and Hakwan Lau. 2013. "Anatomical Coupling Between Distinct Metacognitive Systems for Memory and Visual Perception." *The Journal of Neuroscience* 33 (5): 1897–1906. doi:10.1523/JNEUROSCI.1890-12.2013.

Mendoza-Halliday, Diego, and Julio C Martinez-Trujillo. 2017. "Neuronal Population Coding of Perceived and Memorized Visual Features in the Lateral Prefrontal Cortex." *Nature Communications* 8 (May: 1–13. doi:10.1038/ncomms15471.

Michel, Matthias. 2019. "The Mismeasure of Consciousness: a Problem of Coordination for the Perceptual Awareness Scale." *Philosophy of Science* 86 (5): 1239–49.

Michel, Matthias, and Jorge Morales. 2020. "Minority Reports: Consciousness and the Prefrontal Cortex." *Mind & Language 35*: 493-513. doi:10.1111/mila.12264.

Miyamoto, Kentaro, Rieko Setsuie, Takahiro Osada, and Yasushi Miyashita. 2018. "Reversible Silencing of the Frontopolar Cortex Selectively Impairs Metacognitive Judgment on Non- Experience in Primates." *Neuron* 97 (4): 980–86. doi:10.1016/j.neuron.2017.12.040.

Miyamoto, Kentaro, Takahiro Osada, Rieko Setsuie, Masaki Takeda, Keita Tamura, Yusuke Adachi, and Yasushi Miyashita. 2017. "Causal Neural Network of Metamemory for Retrospection in Primates." *Science* 355 (6321): 188–93. doi:10.1126/science.aal0162.

Morales, Jorge, B Odegaard, and Brian Maniscalco. 2022. "The Neural Substrates of Conscious Perception Without Performance Confounds." In *Neuroscience and Philosophy* (eds. De Brigard, F. & Sinnott-Armstrong, W.). MIT Press.

Morales, Jorge, Hakwan Lau, and Stephen M Fleming. 2018. "Domain-General and Domain-Specific Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex." *The Journal of Neuroscience* 38 (14): 3534–46. doi:10.1523/jneurosci.2360-17.2018.

Morales, Jorge, Jeffrey Chiang, and Hakwan Lau. 2015. "Controlling for Performance Capacity Confounds in Neuroimaging Studies of Conscious Awareness." *Neuroscience of Consciousness* 1 (1). doi:10.1093/nc/niv008.

Norman, Elisabeth, and Mark C Price. 2015. "Measuring Consciousness with Confidence Ratings." In *Behavioral Methods in Consciousness Research*, edited by Morten Overgaard, 159–80. Oxford University Press. doi:10.1093/acprof:oso/9780199688890.003.0010.

Odegaard, Brian, Min Yu Chang, Hakwan Lau, and Sing-Hang Cheung. 2018. "Inflation Versus Filling-in: Why We Feel We See More Than We Actually Do in Peripheral Vision." *Philosophical Transactions of the Royal Society B: Biological Sciences* 373 (1755): 20170345–10. doi:10.1098/rstb.2017.0345.

Persaud, Navindra, Matthew Davidson, Brian Maniscalco, Dean Mobbs, Richard E Passingham, Alan Cowey, and Hakwan Lau. 2011. "Awareness-Related Activity in Prefrontal and Parietal Cortices in Blindsight Reflects More Than Superior Visual Performance." *NeuroImage* 58 (2): 605–11. doi:10.1016/j.neuroimage.2011.06.081.

Persaud, Navindra, Peter McLeod, and Alan Cowey. 2007. "Post-Decision Wagering Objectively Measures Awareness." *Nature Neuroscience* 10 (2): 257–61. doi:10.1038/nn1840.

Peters, Megan A K, and Hakwan Lau. 2015. "Human Observers Have Optimal Introspective Access to Perceptual Processes Even for Visually Masked Stimuli." *eLife* 4. doi:10.7554/eLife.09651.

Peters, Megan A K, Tony Ro, and Hakwan Lau. 2016. "Who's Afraid of Response Bias?" *Neuroscience of Consciousness* 1 (1). doi:10.1093/nc/niw001.

Phillips, Ian. 2016. "Consciousness and Criterion: on Block's Case for Unconscious Seeing." *Philosophy and Phenomenological Research* 93 (2): 419–51. doi:10.1111/phpr.12224.

Phillips, Ian, and Jorge Morales. 2020. "The Fundamental Problem with No-Cognition Paradigms" 24 (3): 165–67. doi:10.1016/j.tics.2019.11.010.

Ramsøy, T Z, and M Overgaard. 2004. "Introspection and Subliminal Perception." *Phenomenology and the Cognitive Sciences* 3: 1–23.

Rausch, Manuel, and Michael Zehetleitner. 2016. "Visibility Is Not Equivalent to Confidence in a Low Contrast Orientation Discrimination Task." *Frontiers in Psychology* 7 (e1004519): 47. doi:10.1093/brain/121.1.25.

Robinson, Zack, Corey J Maley, and Gualtiero Piccinini. 2015. "Is Consciousness a Spandrel?" *Journal of the American Philosophical Association* 1 (2): 365–83. doi:10.1017/apa.2014.10.

Rosenthal, David. 2005. *Consciousness and Mind*. New York: Oxford University Press.

Rosenthal, David. 2008. "Consciousness and Its Function." *Neuropsychologia* 46 (3): 829–40. doi:10.1016/j.neuropsychologia.2007.11.012.

Rosenthal, David. 2012. "Higher-Order Awareness, Misrepresentation and Function." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1594): 1424–38. doi:10.1016/j.concog.2009.12.010.

Rosenthal, David. 2015. "Quality Spaces and Sensory Modalities." In *The Nature of Phenomenal Qualities: Sense, Perception, and Consciousness*, edited by Paul Coates and Sam Coleman, 1–40. Oxford: Oxford University Press.

Rosenthal, David. 2019. "Consciousness and Confidence." *Neuropsychologia* 128: 255–65. doi:10.1016/j.neuropsychologia.2018.01.018.

Rouault, M, A McWilliams, Micah Allen, and Stephen M Fleming. 2018. "Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging." *Personality Neuroscience* 1 (1): e17. doi:10.1017/pen.2018.16.

Rounis, Elisabeth, Brian Maniscalco, John C Rothwell, Richard E Passingham, and Hakwan Lau. 2010. "Theta-Burst Transcranial Magnetic Stimulation to the Prefrontal Cortex Impairs Metacognitive Visual Awareness." *Cognitive Neuroscience* 1 (3): 165–75. doi:10.1080/17588921003632529.

Rouy, M. et al. 2022. "Metacognitive Improvement: Disentangling Adaptive Training from Experimental Confounds." *Journal of Experimental Psychology: General, 151*(9): 151(9): 2083–2091. doi:10.1037/xge0001185.

Samaha, Jason, Luca Iemi, and Bradley R Postle. 2017. "Prestimulus Alpha-Band Power Biases Visual Discrimination Confidence, but Not Accuracy." *Consciousness and Cognition* 54 (September): 47–55. doi:10.1016/j.concog.2017.02.005.

Samaha, Jason, Luca Iemi, Saskia Haegens, and Niko A Busch. 2020. "Spontaneous Brain Oscillations and Perceptual Decision-Making" 24 (8): 639–53. doi:10.1016/j.tics.2020.05.004.

Smith, J David, Justin J Couchman, and Michael J Beran. 2014. "Animal Metacognition: a Tale of Two Comparative Psychologies." *Journal of Comparative Psychology,* 128 (2): 115–31. doi:10.1037/a0033105.

Smith, J David, Wendy E Shields, and David A Washburn. 2003. "The Comparative Psychology of Uncertainty Monitoring and Metacognition." *Behavioral and Brain Sciences* 26 (3): 317–73.

Solovey, Guillermo, Guy Gerard Graney, and Hakwan Lau. 2015. "A Decisional Account of Subjective Inflation of Visual Perception at the Periphery." *Attention, Perception, & Psychophysics*, 77:258-271. doi:10.3758/s13414-014-0769-1.

Stolyarova, A, M Rakhshan, E E Hart, T J O'Dell, M A K Peters, Hakwan Lau, A Soltani, and A Izquierdo. 2019. "Contributions of Anterior Cingulate Cortex and Basolateral Amygdala to Decision Confidence and Learning Under Uncertainty." *Nature Communications* 10 (1): 1–14. doi:10.1038/s41467-019-12725-1.

Tye, Michael. 1996. "The Function of Consciousness." *Noûs* 30 (3): 287–305.

van den Berg, Ronald, Ariel Zylberberg, Roozbeh Kiani, Michael N Shadlen, and Daniel M Wolpert. 2016. "Confidence Is the Bridge Between Multi-Stage Decisions." *Current Biology*, 26: 1–12. doi:10.1016/j.cub.2016.10.021.

van Gaal, Simon, Floris P de Lange, and Michael X Cohen. 2012. "The Role of Consciousness in Cognitive Control and Decision Making." *Frontiers in Human Neuroscience* 6 (May) doi:10.3389/fnhum.2012.00121.

van Gaal, Simon, K Richard Ridderinkhof, Johannes J Fahrenfort, H Steven Scholte, and Victor A F Lamme. 2008. "Frontal Cortex Mediates Unconsciously Triggered Inhibitory Control." *The Journal of Neuroscience* 28 (32): 8053–62. doi:10.1523/JNEUROSCI.1278-08.2008.

Wierzchoń, Michał, Dariusz Asanowicz, and Borysław Paulewicz. 2012. "Subjective Measures of Consciousness in Artificial Grammar Learning Task." *Consciousness and Cognition* 21 (3): 1141–53. doi:10.1016/j.concog.2012.05.012.

Zehetleitner, Michael, and Manuel Rausch. 2013. "Being Confident Without Seeing: What Subjective Measures of Visual Consciousness Are About." *Attention, Perception, & Psychophysics* 75 (7): 1406–26. doi:10.3758/s13414-013-0505-2.