How (not) to underestimate unconscious perception

Matthias Michel[1]

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

Author's version. Forthcoming in *Mind & Language*. Please cite the published version.

Abstract: Studying consciousness requires contrasting conscious and unconscious perception. While many studies have reported unconscious perceptual effects, recent work has questioned whether such effects are genuinely unconscious, or whether they are due to weak conscious perception. Some philosophers and psychologists have reacted by denying that there is such a thing as unconscious perception, or by holding that unconscious perception has been previously overestimated. This article has two parts. In the first part, I argue that the most significant attack on unconscious perception commits the criterion content fallacy: the fallacy of interpreting evidence that observers were conscious of *something* as evidence that they were conscious of the *task-relevant features* of the stimuli. In the second part, I contend that the criterion content fallacy is prevalent in consciousness research. For this reason, I hold that if unconscious perception exists, scientists studying consciousness could routinely underestimate it. I conclude with methodological recommendations for moving the debate forward.

Unconscious perception is like Sisyphus' rock. Each time researchers believe they have proved it, the rock rolls down, and they have to start over again (Irvine, 2012a; Michel, 2020). Tired of having to imagine Sisyphus happy, researchers have developed new tasks with the potential to demonstrate the existence of unconscious perception once and for all (Peters & Lau, 2015). And as these experiments have failed to find a single trace of unconscious perception, the top of the hill looks unattainable. Consciousness researchers are not like Sisyphus, the skeptics argue. They are like Tantalus: desperately reaching for something they cannot have. This article is a response to the skeptics. Unconscious perception is within our grasp.

Determining whether unconscious perception exists is crucial for current discussions on consciousness and perception. Theories of consciousness typically aim to identify the mechanisms distinguishing conscious from unconscious perception (e.g., Brown et al. 2019; Mashour et al. 2020; Lamme, 2015). If unconscious perception does not exist, such theories are probably wrong. The issue is also relevant for philosophical theories of perception: unconscious perception is commonly thought to support 'representationalist' views of perception over 'naïve realist' views (Berger & Nanay, 2016).

The current orthodoxy is that unconscious perception exists. The evidence for this comes from a wide variety of experimental paradigms in which healthy participants – and neurological patients – can identify properties of stimuli that they fail to report perceiving (Breitmeyer, 2015; Kim & Blake, 2005; LeDoux et al. 2020; Weiskrantz, 2009).

Philosophers and psychologists have challenged this orthodox view (e.g., Peters & Lau, 2015; Phillips, 2016). Unconscious perception deniers argue that purported cases of unconscious perception could be interpreted instead as cases of weakly conscious, non-reported perception. In this view, alleged evidence for unconscious perception stems from the fact that subjects often fail to report seeing weakly conscious stimuli: their subjective reports are biased. I will say more about the concept of 'report bias' in what follows. For now, the key point is that when subjects are not sure whether they saw a stimulus or not, as often happens for very weak stimuli, they tend to give a response indicating not seeing anything. If subjects were not biased in this way, the

skeptics argue, evidence for unconscious perception would vanish (Balsdon & Clifford, 2018; Cheesman & Merikle, 1986; Eriksen, 1960; Goldiamond, 1958; Irvine, 2012b, 2019; Merikle, 1982, 1984; Phillips, 2016, 2018, 2021; Reingold & Merikle, 1988, 1990).

To support their claim, the skeptics rely on experimental paradigms controlling for the effects of report biases (more on this in Section 1) (Barthelme and Mamassian, 2009; de Gardelle and Mamassian, 2014; Peters & Lau, 2015). The bad news for unconscious perception enthusiasts is that once biases are controlled for, as done in a series of studies by Peters and colleagues, subjects do not exhibit behaviors consistent with unconscious perception (Knotts et al. 2018; Peters et al. 2017; Peters & Lau, 2015). Following Berger & Mylopoulos (2019, p.1), this series of studies can be considered as "the most powerful evidence" in support of the conclusion that purported cases of unconscious perception are actually cases of weak conscious perception.

If the skeptics are correct, a wide variety of methods usually thought to elicit unconscious perception actually do not (Breitmeyer, 2015; Kim & Blake, 2007). And if this is true, Sisyphus' rock is indeed pushed way downhill. Many experiments carried out by consciousness researchers in the last four decades would pretty much count for nothing. Experimenters thought that they were contrasting conscious and unconscious perception when they were merely contrasting reported and unreported conscious perception. At the very least, unconscious perception would have been significantly overestimated.

Before we move on, it is important to note that this debate is about the existence of unconscious *perception*, not mere unconscious *sensory processing*. For instance, neurons in the primary visual cortex could carry information about the shadows cast by retinal blood vessels (Adams & Horton, 2002). But if that's the case, we are seldom, if ever, conscious of those shadows. Most skeptics would accept unconscious sensory processing, or unconscious sensory registration of this kind (Burge, 2010; Phillips, 2016, 2021). However, they would argue that the mere fact that some neurons respond to this feature is not sufficient to indicate a genuinely perceptual unconscious state. Perception is a personal-level phenomenon: a perceptual state is a state that can be attributed to a subject, not merely to her visual system. Evidence for

unconscious perception would come, for instance, from evidence that an unconscious perceptual state is available for personal-level decisions, such as overt categorization responses.

In this article, I argue that the evidence interpreted as indicating that unconscious perception does not exist is in fact consistent with the hypothesis that there is unconscious perception[1]. My critique of this evidence also applies more generally to a variety of experimental paradigms in consciousness science. I suggest that many procedures used to determine whether subjects are conscious or unconscious of stimuli are too conservative when it comes to attributing unconscious perception. As a result, I claim that if unconscious perception exists, it could be underestimated in consciousness research.

1. Motivating Skepticism

1.1. The criterion problem

Signal Detection Theory is the main framework for interpreting subjects' responses in experimental settings (Green & Swets, 1966; Macmillan & Creelman, 2005). In this framework, a response in a perceptual task is the conjoint product of *sensory evidence*[2] for the presence of a target stimulus, and a *criterion* for responding that the target stimulus is present.

[1] See Berger & Myrtopoulos (2019) for a similar attempt, and Phillips (2020) for a response. Another strategy to respond to the skeptics is to provide cases that meet the criteria for *perception* and *unconsciousness* (e.g., Block, 2016; Quilty-Dunn, 2019; Peters et al., 2017b). While I agree with all these past efforts, my critique extends more generally to the methods used for assessing unconscious perception in the scientific study of consciousness.

[2] While the term 'sensory evidence' is standardly used in Signal Detection Theory (SDT) when it is applied to perceptual decisions, it describes the largely sub-personal process, carried out by the perceptual system, of receiving a signal of a given strength indicating the presence of a given stimulus feature. SDT is not committed to this being 'evidence' in the sense in which *persons* have evidence, or in the sense in which perceptual experience supplies evidence.

For this reason, a subject's observed tendency to answer "not seen" more often than would be expected if she consciously saw the stimulus can have multiple interpretations, depending on the putative *source* of this tendency.

A straightforward explanation is that the subject tends to press the "not seen" button more often than would be expected if she saw the stimuli because she feels like she doesn't see the stimuli. But that's not the only explanation.

One alternative is that the subject has a general preference for pressing a particular response button. For instance, right-handers tend to favor answering with the right index finger. Over the course of a long and difficult experiment, this could bias the subject's responses. This factor should be controlled for.

A more serious worry is that the subject could *misinterpret task instructions*. The instruction to press a button when the stimulus is 'not seen' might have several interpretations, varying from participant to participant. For example, some participants might answer "not seen" when they have a vague feeling of having seen something, but have no idea *what* they saw; while other participants might answer "not seen" only when they have the feeling that nothing was presented on the screen at all. A participant with the former interpretation will tend to press the 'not seen' button more often than a participant with the latter interpretation.

Finally, the observer could be *unable to comply with task instructions*. For example, participants presented with a range of stimulus strengths might be unable to optimize criterion setting for each stimulus strength. Instead, they could set a single criterion across multiple stimulus strengths (Gorea & Sagi, 2000, 2002; Rahnev et al. 2011; Rahnev, 2020; but see Denison et al. 2018). Doing so would usually result in an overly conservative criterion for weak stimuli, since the participant relies on a criterion optimized for mid-range stimulus strength when judging the visibility of weak stimuli.

As most experiments in consciousness research do not control for these last two factors, interpreting the observer's tendency to answer "not seen" as indicating that she does not

consciously perceive the stimuli is not directly justified. Instead, this behavior could be explained by factors that do not have much to do with consciousness, like misinterpretation of task instructions or inability to comply with task instructions.

The *criterion problem*, as it applies in consciousness research, is the problem of determining how the observer's tendency to answer "not seen" should be interpreted. In most experiments, behavior alone underdetermines what interpretation should be favored (Goldiamond, 1958; Eriksen, 1960; Peters et al., 2016; Reingold & Merikle, 1988, 1990). To solve that problem, one should rule out a variety of factors in order to leave the absence of consciousness as the only plausible explanation of the observer's behavior.

1.2. The 2IFC paradigm

One solution to the criterion problem is to use 'bias-free', or at least 'bias-discouraging' tasks, such as 2-interval forced choice paradigms (2IFC). For our purpose, the main study using this paradigm was conducted by Peters & Lau (2015)[3]. In this study, subjects had to discriminate, in two successive intervals, whether a grating was right- or left-tilted. Then, they indicated which of the two discrimination decisions they felt more confident in by betting either on their discrimination decision in the first, or second interval (Figure 1). In a control condition, this betting task was replaced by a task in which subjects reported in which interval the stimulus was more visible. The crucial trick in Peters and Lau's experiment is that, unbeknownst to the subjects, some intervals actually did not contain a target.

[3] In this article, I focus on unconscious perception in healthy subjects. Perhaps the most successful demonstration of unconscious perception to date was achieved by Azzopardi & Cowey (1997) with blindsight patient G.Y., comparing 2IFC detection performance to Yes-No detection performance (see also Yoshida & Isa, 2015; and Michel & Lau (2021) for a discussion of this result). I come back to Azzopardi & Cowey's (1997) experiment in Section 4.

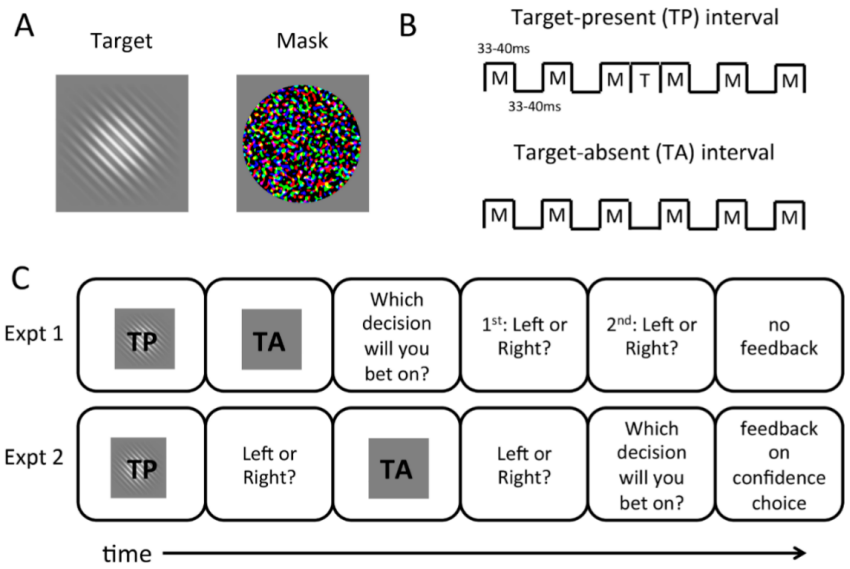

Figure 1. Source: Peters and Lau (2015). 2IFC task. (A) Targets are either right- or left-tilted gratings. (B) Each trial consists of two intervals of discrimination. Some intervals contain a target (TP). In other intervals, the target is replaced by a blank frame (TA). (C) Experimental tasks. In Experiment 1 subjects bet on which discrimination decision they feel more confident in and then indicate the orientation of the gratings in both intervals. In Experiment 2, subjects bet on the interval after the discriminations, and feedback is given.

Now, imagine what happens if you do not see the target consciously. "Seeing" the target feels just like seeing nothing at all. If subjects are not aware of the targets, they should thus judge the target-present intervals to be just as visible as the target-absent intervals. This should translate behaviorally into chance-level performance for responding that the stimulus is more visible in the target-present interval compared to the target-absent interval, or in the betting version of the task, in random bets on either the target-present or target-absent intervals.

So, *consciousness* of the targets can be assessed by analyzing the participants' behavior on the betting task, called the *Type-2* task. Meanwhile, *perception* of the targets is evaluated by analyzing their behavior on the discrimination task, called the *Type-1* task (Galvin et al. 2003). Above-chance performance on the Type-1 task indicates that subjects perceive the targets.

From here, there are two possible outcomes. If subjects are at chance level on the Type-2 task, but above chance on the Type-1 task, they *perceive* the targets *unconsciously*. If subjects are

above chance on the Type-2 task as soon as they are above chance on the Type-1 task, the experiment does not demonstrate unconscious perception.

2IFC paradigms are relatively free from biases (Green and Swets, 1966; Macmillan and Creelman, 2005) and can be used to rule out the different factors mentioned above. First, the target is randomly presented either in the first or second interval: a general tendency to press the button corresponding for instance to 'First Interval' will not move the result in any particular direction. Second, since the observer's judgment is not a matter of 'how much' visibility, but a *relative* judgment, the 2IFC paradigms also provide relatively unambiguous task instructions. Finally, as long as the stimuli presented in the two intervals only vary in one dimension (in this case, the dimension determining target visibility), observers do not have to maintain multiple criteria, or choose how to weight several stimulus dimensions.

Controlling for biases in this way, Peters & Lau (2015) report that subjects do not exhibit behaviors consistent with unconscious perception. When observers can discriminate between stimuli, they also tend to bet on the target-present interval more often than the target-absent interval. The same result is obtained in the control condition in which subjects provide visibility judgments: observers tend to judge that the target is more visible in the target-present compared to the target-absent interval. As such, Peters & Lau's experiment indicates that unconscious perception is not as easy to elicit in healthy subjects as previously thought once criterion biases are controlled for (see also Balsdon & Clifford, 2018).

A weak interpretation of this result, as well as similar results (Peters et al. 2017; Knotts et al. 2018)[4], is that it indicates that unconscious perception has been *overestimated* in most studies in

[4] Several studies have replicated Peters & Lau's result with different suppression techniques, such as visual suppression with transcranial magnetic stimulation (TMS) (Peters et al. 2017), continuous flash suppression, interocular suppression and backward masking (Knotts et al., 2018). However, we should avoid strong interpretations of these results. In the study by Peters et al. (2017) using TMS visual suppression, only about 5% of all trials were actually relevant for assessing unconscious perception – or what the authors call "absolute blindsight". As Koenig & Ro (2019) write: "there was an insufficient number of trials to adequately measure absolute TMS-induced blindsight" (p.219). The study by Knotts et

consciousness research. A stronger interpretation is that most methods thought to elicit unconscious perception actually do not. After controlling for criterion biases, evidence for unconscious perception vanishes (Balsdon & Clifford, 2018; Phillips, 2016, 2018). In the next section, I hold that this strong interpretation is unwarranted. In Section 3, I will argue that, if it exists, unconscious perception could be underestimated in consciousness research.

2. Conscious… of what?

When we say that subjects are conscious or unconscious, we should always ask ourselves: conscious or unconscious... *of what*? It is one thing to be conscious that *something* was presented. It is another to be conscious of the *target stimulus*, and yet another to be conscious of the *task-relevant feature* of the stimulus. I can see the same briefly flashed stimulus as a dog, as an animal, as a bunch of unclear geometrical lines, or as a brief flash of light (Bowers & Jones, 2008; Mack et al. 2008; Mack & Palmeri, 2010; Stazicker, 2011; Straube & Fahle, 2011; Thomas, 1985).

al. (2018) compared a variety of visual suppression techniques to backward and forward masking, and observed that all visual suppression methods equally impacted Type-1 and Type-2 performances. If one accepts Peters & Lau's (2015) original conclusion, this study thus indicates that a wide variety of methods are equally bad at inducing unconscious perception as forward and backward masking. But if one doesn't accept that conclusion, the study indicates that a wide variety of methods are equally *good* at inducing unconscious perception as forward and backward masking. Below, I argue that a strong interpretation of Peters & Lau's result is unwarranted. If so, interpreting Knotts et al.'s result as indicating that a variety of visual suppression techniques actually do not induce unconscious perception is unwarranted as well.

Stimuli have many features like shape, size, color, contrast, or duration[5]. But only some of those features are relevant to perform the Type-1 task successfully. In Peters & Lau's study, the *orientation* of the gratings is the task-relevant feature. That is, participants cannot perform this task above-chance without forming a representation of *that* specific feature.

Asking whether unconscious perception is possible amounts to asking whether *this* representation can determine the discrimination decision without the observer being conscious of its content. For Type-2 judgments to be informative with respect to that question, they must indicate that subjects were *conscious* of the *task-relevant features* of the stimuli — that is, conscious of the very features that they used to perform the Type-1 task successfully.

As Newell & Shanks (2014) argued, assessment of awareness should, as much as possible, target only the information that is relevant for performing the behavioral task (See also Shanks & St John, 1994). If the measures used to infer perception and awareness are not about the same contents, then one is bound to either overestimate or underestimate unconscious processing (Newell & Shanks p.4).

In Peters & Lau's experiment, betting judgments are not determined only — and perhaps not even primarily — by the conscious perception of task-relevant features. Seeing just a vague contour, or a contrast difference between the target and the mask, or having the feeling that something grey-ish was present between the masks, is sufficient to bet on the target present

[5] Task-irrelevant features are represented along with task-relevant features (Marshall & Bays, 2013; Xu, 2010). Importantly, visual masking has different effects on various stimulus features (Breitmeyer, 1984; Kahneman, 1968; Koivisto & Neuvonen, 2020; Ogmen et al., 2003), as well as on detection and discrimination performances (Jimenez et al. 2019). For instance, optimal suppressions of a target's contour and target's brightness with metacontrast and paracontrast masking occur at different asynchronies between the mask and the target (Breitmeyer et al., 2006; Stober et al., 1978). Similar results have been obtained with other suppression techniques, such as binocular rivalry paradigms (Gelbart-Sagiv et al. 2016; Hong & Blake, 2009; Zadbood et al. 2011), or visual crowding (Doerig et al. 2019; Manassi & Whitney, 2018). Hence, one cannot simply assume that all stimulus features, both task-relevant and task-irrelevant, are equally impacted by awareness suppression techniques.

interval above chance. Whereas the Type-1 task requires representing task-relevant features, subjects can correctly bet on the target-present interval without consciously representing any of those features. Having the feeling of seeing *something* in one interval compared to the other, however indeterminate that impression, is sufficient to bet on the target-present interval above chance.

For this reason, based on a subject's above-chance betting performance, one can conclude that the subject consciously perceived that *something* was present in one of the two intervals, or that *something was different* between the two intervals. But nothing warrants the stronger conclusion that the subject bet on the target present interval above chance *because* she was conscious of the very features that led her to perform the Type-1 task above chance.

This leaves an alternative interpretation open. Although observers subjectively distinguish the target-present from the target-absent interval, since they see that *something* is present in one of the two intervals, they successfully perform the Type-1 task based on an *unconscious* representation of the *task-relevant feature* of the stimulus[6].

To put it bluntly, subjects feel like they see a grey-ish blob for a very brief moment in one interval, and not the other. Because seeing *something* is better than seeing *nothing at all*, they bet on the interval in which they saw something. But while they *feel* like they see a grey-ish blob, subjects *unconsciously* perceive the orientation of the grating. This unconscious perception of the task-relevant feature drives subjects' Type-1 performance, while the conscious perception of *something* concomitantly increases their betting performance The unconscious perception of the

[6] By accepting this interpretation of the evidence, one is committed to a weak version of the 'partial awareness hypothesis' (Kouider et al. 2010), according to which different levels of representations can be consciously accessed independently. This version of the partial awareness hypothesis only requires one to accept that one can be conscious of the surface features of a stimulus (e.g. the color of the stimulus) without being conscious of its contour features (e.g. its orientation) (see Breitmeyer (2014) for a book-length defense of this view, partly based on a review of the evidence mentioned in Footnote 4). For the purpose of this article, I remain neutral on what other kinds of dissociations could be observed, for example on the question of whether high-level features such as object category can be consciously accessed independently of low-level features, or whether contour features (e.g. shape) could be consciously experienced without any experience of surface features (e.g. color).

orientation of the stimulus is hidden within the conscious perception of a grey blob – an instance of unconscious perception within conscious perception (See Block's contribution to Peters et al. 2017b).

Peters & Lau (2015) themselves suggested a similar interpretation for their result (see Section '2IFC detection'). However, they discarded it on the basis that such an interpretation cannot account for the 'flat' Type-2 false alarm rate observed in the data. The Type-2 false alarm is defined by Peters & Lau as incorrect orientation discrimination and bet on target-present interval. Suppose subjects were really performing the betting task simply by consciously detecting the presence of *something*. In that case, we should expect the Type-2 false alarm rate to increase as signal strength increases, since subjects are more likely to bet on the target-present interval with increasing signal strength. This, they argue, is incompatible with the observed flat Type-2 false alarm rate. However, as Phillips (2021) noted, subjects are also more likely to perform the discrimination task correctly as signal strength increases, which should in turn *decrease* the participants' Type-2 false alarm rate (as defined above). A flat Type-2 false alarm rate could thus result from these two countervailing influences.

To be clear, Peters & Lau do provide an answer to the question of knowing whether, in their visual masking task, totally unconscious percepts can represent orientation. The answer is no. But this is the right answer to the wrong question. The right question is whether there can be an unconscious representation of orientation, even if that representation is within a conscious percept. That is, one wants to assess whether subjects are conscious of the features that drive their Type-1 discrimination performance, not whether they are conscious of *anything* at all. And Peters & Lau do not answer that question. So, their result does *not* demonstrate that subjects need to be conscious *of the orientation* if they are to perform above-chance on the discrimination task. It merely shows that they need to be conscious *of something*. But being conscious of something is compatible with unconsciously representing the task-relevant features of the stimuli. And if that's the case, this experiment is far from demonstrating that unconscious perception does not exist.

Before explaining how the same reasoning applies to various procedures for detecting consciousness, let me be clear on what I do, and what I do not claim. I do not claim that scientists should not use 2IFC paradigms. Under some conditions, a *positive* result with a 2IFC paradigm would suggest unconscious perception (See section 4). However, a *null result* with a 2IFC paradigm – like all null results – should be interpreted carefully. A failure to meet the high-standard set by the 2IFC approach does not mean that unconscious perception does not exist, or that all previous unconscious effects demonstrated in the literature resulted from uncontrolled response criteria.

3. The criterion content fallacy

Research on consciousness has often ignored what Kahneman (1968) called the *criterion content*: the set of features on which observers base their perceptual (or metacognitive) decisions. Differences in the criterion contents used to perform the Type-1 and Type-2 tasks often go uncontrolled: Type-1 and Type-2 judgments are not based on the same features. Type-2 judgments usually have a permissive criterion content: they can be based on the perception of *any* visual features (or even non-visual features, as I explain in Section 3.2). Type-1 judgments in discrimination tasks typically have a more restrictive criterion content: these judgments are based on the perception of task-relevant (visual) features.

The discrepancy between permissive and restrictive criterion contents leads researchers to commit the *criterion content fallacy*: the fallacy of concluding that subjects are conscious of the *task-relevant features* of the stimuli based on the mere evidence that they are conscious of *something*. Committing this fallacy, in turn, leads researchers to systematically overestimate the observers' consciousness of the visual features used to perform the Type-1 discrimination task. Trials in which the subjects' reports do not warrant the conclusion that they are conscious of the task-relevant features of the stimuli are nevertheless categorized as conscious trials. To illustrate this fallacy, let me focus on procedures relying on the Perceptual Awareness Scale (PAS).

3.1. A case study: the Perceptual Awareness Scale

The PAS allows subjects to provide judgments with the following response categories: "No experience", "Brief glimpse", "Almost clear image", "Absolutely clear image" (Ramsøy & Overgaard, 2004; Sandberg et al. 2010). It has been used in a wide variety of studies. Here's a small, non-exhaustive sample: the PAS is involved in determining whether working memory contents can be represented unconsciously (Trübutschek et al. 2019), whether blindsight subjects have weak conscious experiences (Overgaard et al. 2008; Mazzi et al. 2016), whether conscious perception is graded or all-or-none (Windey et al. 2014), whether subjects have weak conscious experiences in exclusion tasks (Sandberg et al. 2014), as well as to identify neural correlates of consciousness (e.g., Andersen et al. 2016; Tagliabue et al. 2016).

Despite its widespread use, remarkably little work has been done to investigate the validity of the interpretation of PAS ratings as indications of conscious perception (Michel, 2019) [7]. If using the PAS leads researchers to commit the criterion content fallacy, unconscious perception of task-relevant features could have been underestimated in all those studies – among many others. I believe that's the case. Let me explain why.

Experimenters typically describe the second level on the PAS – usually called "Brief glimpse" – as "A feeling that something has been shown. Not characterized by any content and this cannot be specified any further" (Ramsøy & Overgaard, 2004, p.12). That is, subjects use the

[7] It is not clear whether tasks relying on the PAS should count as Type-2 tasks or not. Galvin et al. (2003) defined a Type-1 task as a task in which "an observer decides which of two or more events *defined independently of the observer* has occurred" (p.843), while they restrict Type-2 tasks to "The task of discriminating between one's own correct and incorrect Type 1 decisions" (p.843). The PAS was introduced as a scale for judging the 'clarity' *of one's own conscious experience* of a stimulus (Ramsøy & Overgaard, 2004). This dimension is not 'defined independently of the observer'. On the other hand, it is possible that the subjects simply rate the perceived clarity *of the stimulus* (Michel, 2019), essentially committing what Titchener called "the stimulus error" (Tichener, 1905; Boring, 1921). In this latter case, the observers' judgments are about objectively definable properties, such as, for instance, stimulus duration and contrast.

"Brief glimpse" response category when they see *something*, but do not consciously perceive the *task-relevant features* of the stimuli.

With a few exceptions (e.g. Melloni et al. 2011), most researchers interpret trials in which subjects report seeing just a "brief glimpse" as conscious trials — trials in which the sensory information used to perform the Type-1 task was conscious (e.g. Bergström & Eriksson, 2014; King et al. 2016; Soto et al. 2011; Trübutschek et al. 2019). Yet, consciously seeing *something*, or just a 'brief glimpse', is compatible with *unconsciously* representing *task-relevant features*. And it is even compatible with the complete absence of perception of the task-relevant features.

Indeed, in a study by Mazzi et al. (2016), subjects often used the "brief glimpse" report category even when they were at chance on the Type-1 discrimination task, indicating no perception of the task-relevant features. Similarly, in a study by Jimenez et al. (2019) participants performed a detection task followed by a discrimination task, and then rated perceptual awareness with the PAS. Subjects often reported "weak perception" even if they couldn't perform the discrimination task above chance, but could only detect the presence of a stimulus. These studies indicate that the "brief glimpse" report category is not tied to the conscious representation of task-relevant features. If so, it should not be interpreted as indicating conscious perception of those features.

Doing otherwise amounts to committing the criterion content fallacy. Given the way in which the report category is described to the subjects, the use of the "brief glimpse" report category does not warrant the conclusion that the subject was conscious of the *task-relevant feature* of the stimulus (Dienes & Seth, 2010; Jimenez et al. 2019). Instead, one can only conclude that the subject saw *something*. And again, *this* weaker conclusion is compatible with the *unconscious* representation of the task-relevant feature of the stimulus *within* the (indeterminate) conscious percept. As a result, interpreting trials in which participants report seeing just a "brief glimpse" as conscious trials could lead researchers to systematically underestimate unconscious perception of task-relevant features. Cases in which subjects unconsciously represent the task-relevant

features of the stimuli and successfully perform the Type-1 task are nevertheless counted as conscious trials.

3.2. What about visibility ratings and confidence ratings?

The same reasoning applies to ordinary visibility ratings – i.e., visibility ratings that do not rely on the PAS categories but only on "seen" or "not seen" reports – although perhaps to a lesser extent (e.g., Sergent & Dehaene, 2004).

It is not always clear what "seen" or "not seen" reports are *about*. Do they indicate that the subject saw *the target stimulus*? Or a given *feature* of that stimulus? Or just that they saw *something*? Different experiments are likely to induce different criterion contents depending on how experimenters introduce the response categories to the subjects. Given the ambiguity of the report categories, subjects probably often come up with their own criterion content.

For this reason, it is not clear how scientists should interpret visibility ratings. If the goal is to determine whether subjects are conscious or unconscious of the features of the stimuli that they use to perform the Type-1 task, it should be made explicit to them that their reports should be about the visibility of the task-relevant features of the stimuli. And while this recommendation could help reduce the problem to some extent, it is far from being sufficient. As Galvin et al. (2003) recognized, even with more explicit task instructions, "It is an unavoidable aspect of psychological experiments that no matter how well instructed, it is the observer who determines how the task is done" (p.862).

Scientists also use confidence ratings as indications of consciousness (Norman & Price, 2015). Procedures based on confidence ratings seem to fare a bit better than other procedures with respect to the criterion content fallacy. Across several studies (Rausch et al., 2015, 2018; Rausch & Zehetleitner, 2016; Zehetleitner & Rausch, 2013), Rausch & Zehetleitner have observed that confidence judgments seem to be more influenced by the subjects' conscious perception of the task-relevant features of the stimuli than visibility judgments. For our purposes,

an important result is that while confidence ratings tend to track the accuracy of the decision, visibility ratings tend to track the objective strength of the stimulus (e.g. contrast) even when the participants' responses are incorrect (Rausch & Zehetleitner 2016).

Instead of going into the details of these experiments, let me just explain why this result makes intuitive sense. Task-relevant and task-irrelevant features both contribute to the overall visibility of the stimulus. On the other hand, one's confidence that one has performed the Type-1 task correctly should be particularly sensitive to the perception of the very features that one used to perform the Type-1 task. If the only thing you saw is entirely irrelevant for the task, that shouldn't make you more confident that your response was correct. Since we should expect confidence rating procedures to be more sensitive to the subjects' conscious perception of the task-relevant features of the stimuli than other methods, using confidence ratings as indications of consciousness might fare better than other methods when it comes to mitigating the effects of the criterion content fallacy.

However, confidence judgments are not a perfect solution to the criterion content fallacy. As noted by Rausch et al. (2018, forthcoming), although task-irrelevant features are probably weighted differently for confidence and visibility judgments, one should still expect confidence judgments to be somewhat influenced by the conscious perception of task-irrelevant features. That's because the strength of the sensory evidence for task-irrelevant features is a good proxy for the strength of sensory evidence for task-relevant features, which is in turn a good indication of Type-1 performance.

In addition, confidence judgments could be influenced by additional sources of information that are not available for the Type-1 decision (Berger & Mylopoulos, 2019; Fleming & Daw, 2017; Rosenthal, 2018). First, sensory evidence continues to accumulate after the Type-1 decision is reached (Moran et al. 2015; Murphy et al. 2015; Pleskac & Busemeyer, 2010). So, in some cases there could be more sensory evidence at the time of the Type-2 decision than for the Type-1 decision alone, which inflates Type-2 sensitivity – how efficiently the observer's confidence ratings discriminate correct from incorrect decisions (Balsdon et al. 2020; Charles et

al. 2013; Charles & Yeung, 2019; Rabbit & Yvas, 1981; Resulaj et al. 2009). Second, the act of reporting one's decision on the Type-1 task could be used as evidence for confidence judgments (Fleming & Daw, 2017). For instance, Fleming et al. (2015) showed that changing the motor fluency of the subjects' reports by applying TMS to the motor cortex during their Type-1 decision reports could either increase or decrease their subsequent confidence judgments (see also Gajdos et al. 2019; Kvam et al. 2016; Siedlecka et al. 2016; Wierzchoń et al. 2014). With those additional sources of evidence influencing confidence judgments, those judgments will tend to track the correctness of the Type-1 decisions more accurately than they would otherwise. And as these additional sources of evidence are unavailable for the Type-1 decision, the sensitivity of confidence rating procedures could be inflated by cues contributing to the Type-2 – but not Type-1 – decision, thus leading researchers to overestimate conscious perception.

In most experiments, however, the effect of incorporating additional information to the Type-2 decision could be out-weighted by additional noise affecting the Type-2 decision process (Shekhar & Rahnev, 2020). That is, noise affecting the Type-2 decision process could reduce Type-2 sensitivity, thus potentially countervailing the increase in sensitivity afforded by additional sources of evidence at the Type-2 decision level. For this reason, some dissociations between Type-1 and Type-2 performances can be expected in nearly all experiments. The question is *how much* of a dissociation would be enough to demonstrate unconscious perception. I now attempt to address this point with some recommendations on how to move the debate forward.

4. Moving forward: Model-based assessment of unconscious perception

Peters & Lau's (2015) experiment is on the right track. It can be improved by striving to minimize unnecessary differences in task-irrelevant features between the target-absent and target-present intervals, such as, for instance, differences in luminance. In line with Peters & Lau's paradigm, unconscious perception would thus be indicated by above-chance performance on a Type-1 task discrimination task while Type-2 performance is at chance-level. This result

would provide good evidence that, while subjects do perceive the relevant feature, perceiving it feels just like not perceiving it.

But at this point we face two problems. The first is to determine whether Type-2 performance really is 'at chance'. It's easy enough to prove that a coin is unfair; but how many tosses does it take to prove that it is completely fair? A similar issue arises for identifying 'chance-level' Type-2 performance (Dienes, 2015).

The second problem is to determine how much of a discrepancy between Type-1 and Type-2 performances would count as indicating unconscious perception. After all, we already saw that Type-2 performance should generally be expected to be somewhat suboptimal given that additional noise affects the Type-2 decision process.

A study by Azzopardi & Cowey (1997) with blindsight patient G.Y. shows us the way to solve these two problems at once. Blindsight is typically characterized by low sensitivity in Yes/No (YN) detection tasks – a measure of the capacity to determine whether the stimulus was present or absent. Blindsight patients tend to report not seeing the targets. However, sensitivity can be pretty high in two-alternatives forced-choice (2AFC) detection tasks in which the patient is forced to identify in which interval a stimulus was presented (Weiskrantz, 2009).

Does high 2AFC performance based on stimuli that the patient denies seeing in a YN detection task indicate unconscious perception? As opponents to unconscious perception in blindsight are keen to point out, this discrepancy can alternatively be explained by the fact that blindsight patients are simply reluctant to report degraded conscious experiences, thus adopting an overly conservative criterion in the YN detection task (Campion et al. 1983; Phillips, 2021; but see Michel & Lau, 2021).

Now comes the beauty of Azzopardi & Cowey's (1997) experiment. SDT specifies the mathematical relation between these two tasks (Macmillan & Creelman, 2005, p.168; but see Yeshurun et al. 2008). In particular, sensitivity for a 2AFC task is related to sensitivity for a corresponding YN detection task by a factor of $\sqrt{2}$. This relation can be exploited to get a

non-biased estimate of YN detection sensitivity (YN detection *d'*). If the difference in sensitivity between the two tasks is best explained by response bias affecting the YN detection task only, once corrected by a factor of $\sqrt{2}$, YN detection *d'* should be equal to 2AFC detection *d'*. If, on the other hand, response bias is insufficient to explain the difference in sensitivity between the two tasks, 2AFC detection *d'* should remain higher than the corrected YN detection *d'*. Azzopardi & Cowey (1997) obtained the latter result with patient G.Y. (later confirmed in monkeys by Yoshida & Isa, 2015):

> Because, empirically, 2AFC detection tasks yield identical values of sensitivity as do YN tasks in normal subjects, once scaled by a factor of $1/\sqrt{2}$ in accordance with SDT (as confirmed with control subjects under the present conditions), it follows that the hemianopic patient's residual vision is unlike normal, near-threshold vision and that his brain processes information about the visual stimulus in his scotoma in an unusual way. (p.14193)

I suggest that the same general idea can be preserved for testing unconscious perception in healthy subjects. For a given task, one can start by mapping the relation between Type-1 and Type-2 performances in conditions in which the visibility of the stimuli is left unaltered. Once the relation between those two tasks is well understood, one can mathematically estimate, for any given Type-1 performance, the expected Type-2 performance if subjects were conscious of all the sensory information used to perform the Type-1 task[8]. This model of the subject's ideal Type-2 performance can then be used as a benchmark against which to judge patterns of Type-1 & Type-2 performance in experimental conditions aimed at manipulating the visibility of the stimuli.

This approach combined with a non-biased paradigm solves our problems. First, unconscious perception is not indicated by 'chance-level' Type-2 performance. Instead, it is indicated by the breakdown of the relation between Type-1 and Type-2 performances that would have been expected if subjects were conscious of all the sensory information available to

[8] This model-based approach is already the main idea behind the commonly used indicator *meta-d'* (Maniscalco & Lau, 2012).

perform the Type-1 task (See Balsdon & Clifford, 2018 for a similar approach). Second, this approach can be used to estimate how much of a discrepancy between Type-1 and Type-2 performances would indicate unconscious perception: it is the level of discrepancy at which a model postulating ideal access to the sensory information used for the Type-1 task doesn't provide a satisfying fit to the data (compared to alternative, sub-ideal models). Third, since this approach is assumed to rely on 2IFC tasks, Type-2 bias doesn't provide a good alternative explanation for the observed discrepancy.

Ultimately, the goal is to leave unconscious perception as the only remaining explanation for the fact that subjects can perform a Type-1 task well above the level that would be expected given their (relative lack of) capacity to identify in which of two intervals a target feature was presented. The strongest form of evidence for unconscious perception would thus come from behaviors exhibiting a significant deviation from ideal Type-2 performance – indicating an unconscious effect, in conditions where Type-1 performance is well above-chance – indicating that subjects indeed *perceived* the stimuli.

Conclusion

I have argued that the main study purporting to show that unconscious perception is in fact weak conscious perception is consistent with an interpretation in terms of unconscious perception (Peters & Lau, 2015). While it is too early to know whether unconscious perception exists in healthy subjects or not, I have argued that if unconscious perception exists, it could be routinely underestimated in studies relying on the PAS, or visibility ratings. I have also suggested an approach to move the debate forward.

Nevertheless, the accuracy and validity of detection procedures partly depend on what one wants to do with them. Being clear on *what* detection procedures can validly detect should allow experimenters to identify the best method given their current goals (Michel, 2019). For instance, as they are currently designed, 2IFC paradigms are perfectly adequate if one's goal is to

make sure that subjects do not see anything at all. Here, I aimed to argue that this high-standard is not necessary for demonstrating unconscious perception and that, accordingly, unconscious perception could be within our grasp after all.

Acknowledgments

I thank Ned Block, Axel Cleeremans, Hakwan Lau, Jorge Morales, and two anonymous reviewers for their comments on this paper.

References

Andersen, L. M., Pedersen, M. N., Sandberg, K., and Overgaard, M. (2016). Occipital MEG Activity in the Early Time Range (<300 ms) Predicts Graded Changes in Perceptual Consciousness. Cerebral Cortex, 26(6):2677–2688.

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

Barthelme, Simon and Mamassian, Pascal (2009). Evaluation of objective uncertainty in the visual system. *PLOS Computational Biology*, 5(9):1–8.

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

Balsdon, Tarryn, Wyart, Valentin, & Mamassian, Pascal (2020). Confidence controls perceptual evidence accumulation. *Nature Communications*, *7*(2020).

Berger, Jacob, & Mylopoulos, Myrto (2019). On Skepticism about Unconscious Perception. *Journal of Consciousness Studies*, 1–21.

Berger, Jacob, & Nanay, Bence (2016). Relationalism and unconscious perception. *Analysis*, *76*(4), 426–433.

Bergström, F., & Eriksson, J. (2014). Maintenance of non-consciously presented information engages the prefrontal cortex. *Frontiers in Human Neuroscience*, *8*(November), 938.

Boring, E. G. (1921). The Stimulus-Error. *The American Journal of Psychology*, *32*(4), 449–471.

Block, Ned (2016). The Anna Karenina Principle and Skepticism about Unconscious Perception. *Philosophy and Phenomenological Research*, *93*(2), 452–459.

Bowers, Jeffrey S., & Jones, Keely W. (2008). Detecting objects is easier than categorizing them. *Quarterly Journal of Experimental Psychology*, *61*(4), 552–557.

Breitmeyer, Bruno G., Kafaligönül, Hulusi, Öğmen, Haluk, Mardon, Lynn, Todd, Steven, Ziegler, Ralph (2006). Meta- and paracontrast reveal differences between contour- and brightness-processing mechanisms. *Vision Research*, 46(17), 2645–2658.

Breitmeyer, B. G. (2014) *The Visual (Un)Conscious and Its (Dis)Contents: A microtemporal approach*. Oxford: Oxford University Press.

Breitmeyer, B. G. (2015). Psychophysical "blinding" methods reveal a functional hierarchy of unconscious visual processing. *Consciousness and Cognition*, *35*, 234–250.

Brown, Richard, Lau, Hakwan, & LeDoux, Joseph E. (2019). Understanding the Higher-Order Approach to Consciousness. *Trends in Cognitive Sciences*, *23*(9), 754–768.

Charles, Lucie, Van Opstal, Filip, Marti, Sébastien, & Dehaene, Stanislas (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage*, 73, 80–94.

Charles, Lucie, & Yeung, Nick (2019). Dynamic sources of evidence supporting confidence judgments and error detection. *Journal of Experimental Psychology:Human Perception and Performance*. 45(1), 39–52.

Cheesman, Jim, & Merikle, Philip M. (1986). Distinguishing Conscious from Unconscious Perceptual Processes. *Canadian Journal of Psychology 40*(4), 343–367.

de Gardelle, Vincent and Mamassian, Pascal (2014). Does confidence use a common currency across two visual tasks? *Psychological Science*, 25(6):1286–1288.

Denison, R. N., Adler, W. T., Carrasco, M., & Ma, W. J. (2018). Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences*, 115(43), 11090–11095.

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

Dienes, Z (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In M. Overgaard (Ed.), *Behavioural Methods in Consciousness Research*. Oxford: Oxford University Press, pp 199-220.

Doerig, Adrien, Bornet, Alban, Rosenholtz, Ruth, Francis, Gregory, Clarke, Aaron M., & Herzog, Michael H. (2019). Beyond Bouma's window: How to explain global aspects of crowding? *PLoS Computational Biology*, 15(5), 1–28.

Eriksen, Charles W. (1960). Discrimination and learning without awareness: A methodological survey and evaluation. *Psychological Review*, *67*(5), 279–300.

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

Fleming, Steven M., Maniscalco, Brian, Ko, Yoshiaki, Amendi, Namena, Ro, Tony, & Lau, Hakwan (2015). Action-specific disruption of perceptual confidence. *Psychological Science*, 26, 89–98.

Gajdos, Thibault, Fleming, Steven M., Saez Garcia, Marta, Weindel, Gabriel, & Davranche, Karen (2019). Revealing subthreshold motor contributions to perceptual confidence. *Neuroscience of Consciousness*, *2019*(1), 1–8.

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

Gelbard-Sagiv, Hagar, Faivre, Nathan, Mudrik, Liad, & Koch, Christof (2016). Low-level awareness accompanies "unconscious" high-level processing during continuous flash suppression. *Journal of Vision*, *16*(1), 1–16.

Goldiamond, Israel (1958). Indicators of perception: I. Subliminal perception, subception, unconscious perception: An analysis in terms of psychophysical indicator methodology. *Psychological Bulletin*, *55*(6), 373–411.

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.

Gorea, A., & Sagi, D. (2002). Natural extinction: A criterion shift phenomenon. Visual Cognition, 9(8), 913–936.

Green, David M., & Swets, John A. (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York.

Hillis, James M., & Brainard, David H. (2007). Distinct Mechanisms Mediate Visual Detection and Identification. *Current Biology*, *17*(19), 1714–1719.

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

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

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

Irvine, Elizabeth (2019). Developing Dark Pessimism Towards the Justificatory Role of Introspective Reports. *Erkenntnis*, (0123456789).

Jimenez, Mikel, Villalba-García, Cristina, Luna, Dolores, Hinojosa, José Antonio, Montoro, Pedro R. (2019). The nature of visual awareness at stimulus energy and feature levels: A backward masking study. *Attention, Perception, and Psychophysics*, 81(6), 1926–1943.

Kahneman, Daniel (1968). Method, findings, and theory is studies of visual masking. *Psychological Bulletin*, 70, 404–425.

Kim, Chan Youn, & Blake, Randolph (2005). Psychophysical magic: Rendering the visible "invisible." *Trends in Cognitive Sciences*, *9*(8), 381–388.

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

Koenig, Luca, & Ro, Tony (2019). Dissociations of conscious and unconscious perception in TMS-induced blindsight. *Neuropsychologia*, *128*(March 2018), 215–222.

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

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

Kvam, Peter D., Pleskac, Timothy J., Yu, Shuli, & Busemeyer, Jerome R. (2015). Interference effects of choice on confidence: Quantum characteristics of evidence accumulation. *PNAS*, 112, 10645–10650.

Lamme, Victor (2015). The Crack of Dawn. *Open MIND* (Vol. 22).

Mack, Michael L., Gauthier, Isabel, Sadr, Javid, & Palmeri, Thomas J. (2008). Object detection and basic-level categorization: Sometimes you know it is there before you know what it is. *Psychonomic Bulletin and Review*, *15*(1), 28–35.

Mack, Michael L., & Palmeri, Thomas J. (2010). Decoupling object detection and categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1067–1079.

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

Marshall, Louise, & Bays, Paul (2013). Obligatory encoding of task-irrelevant features depletes working memory resources. *Journal of Vision*, 12(9), 853.

Manassi, Mauro, & Whitney, David (2018). Multi-level Crowding and the Paradox of Object Recognition in Clutter. *Current Biology*, 28(3), R127–R133.

Mashour, George A., Roelfsema, Pieter, Changeux, Jean-Pierre, & Dehaene, Stanislas (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776–798.

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

Melloni, L., Schwiedrzik, C. M., Muller, N., Rodriguez, E., & Singer, W. (2011). Expectations Change the Signatures and Timing of Electrophysiological Correlates of Perceptual Awareness. *Journal of Neuroscience*, *31*(4), 1386–1396.

Merikle, Philip M. (1982). Unconscious perception revisited. *Perception & Psychophysics*, *31*(3), 298–301.

Merikle, Philip M. (1984). Toward a definition of awareness. *Bulletin of the Psychonomic Society*, *22*(5), 449–450.

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

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

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

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

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

Ogmen, Haluk, Breitmeyer, Bruno G., & Melvin, Reginald (2003). The what and where in visual masking. *Vision Research*, 43(12), 1337–1350.

Overgaard, M., Fehl, K., Mouridsen, K., Bergholt, B., & Cleeremans, A. (2008). Seeing without seeing? Degraded conscious vision in a blindsight patient. *PLoS ONE*, *3*(8), 8–11.

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

Peters, Megan A. K., Ro, Tony, & Lau, Hakwan (2016). Who's afraid of response bias? *Neuroscience of Consciousness*, *2016*(1), 1–16.

Peters, Megan A. K., Fesi, Jeremy, Amendi, Namema, Knotts, JD, Lau, Hakwan, & Ro, Tony (2017a). Transcranial magnetic stimulation to the visual cortex induces suboptimal introspection. *Cortex*, *97*, 119–132.

Peters, Megan A. K., Kentridge, Robert W., Phillips, Ian, & Block, Ned (2017b). Does unconscious perception really exist? Continuing the ASSC20 debate. *Neuroscience of Consciousness*, *3*(1), 1–11.

Phillips, Ian (2016). Consciousness and criterion: On block's case for unconscious seeing. *Philosophy and Phenomenological Research*, 93(2):419–451.

Phillips, Ian (2018). Unconscious perception reconsidered. *Analytic Philosophy*, 59(4):471–514.

Phillips, I. (2021). Scepticism about unconscious perception is the default hypothesis. *Journal of Consciousness Studies*, *28*(3–4), 186–205.

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

Quilty-Dunn, Jake (2019). Unconscious perception and phenomenal coherence. *Analysis*, *79*(3), 461–469.

Rabbitt, Patrick, & Vyas, Subhash (1981). Processing a display even after you make a response to it. how perceptual errors can be corrected. *The Quarterly Journal of Experimental Psychology Section A*, *33*(3), 223–239.

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.

Rahnev, D. (2021). Confidence-accuracy dissociation via criterion attraction. *PsyArXiv*. 10.31234/osf.io/5p6x9

Rausch, Manuel, Hellmann, Sebastian, and Zehetleitner, Michael (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, perception & psychophysics*, 80(1):134–154.

Rausch, Manuel, Müller, Hermann J., and Zehetleitner, Michael (2015). Metacognitive sensitivity of subjective reports of decisional confidence and visual experience. *Consciousness and Cognition*, 35:192–205.

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

Rausch, Manuel and Zehetleitner, Michael (forthcoming). Cognitive modelling reveals that visibility judgments rely on two dimensions of sensory evidence. Preprint.

Resulaj, Arbora, Kiani, Roozbeh, Wolpert, Daniel M., & Shadlen, Michael N. (2009). Changes of mind in decision-making. *Nature*, 461(7261), 263–266.

Rosenthal, David (2018). Consciousness and confidence. *Neuropsychologia*, 128:255-265.

Reingold, Eyal M. and Merikle, Philip M. (1988). Using Direct and Indirect Measures to Study Perception without Awareness, *Perception and Psychophysics*, 44: 563–75.

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

Sandberg, Kristian, Timmermans, Bert, Overgaard, Morten, & Cleeremans, Axel (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*, 19(4), 1069–1078.

Sandberg, K., Del Pin, S. H., Bibby, B. M., & Overgaard, M. (2014). Evidence of weak conscious experiences in the exclusion task. *Frontiers in Psychology*, *5*(SEP), 1–8.

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

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

Siedlecka, Marta, Paulewicz, Borysław, & Wierzchon , Michał (2016). But I was so sure! Metacognitive judgments are less accurate given prospectively than retrospectively. *Frontiers in Psychology*, 7(February).

Soto, D., Mäntylä, T., & Silvanto, J. (2011). Working memory without consciousness. *Current Biology*, *21*(22), R912–R913.

Stazicker, James (2011). Attention, Visual Consciousness and Indeterminacy. *Mind and Language*, *26*(2), 156–184.

Stober, R. Stephen, Brussel, Edward M., & Komoda, Melvin K. (1978). Differential effects of metacontrast on target brightness and clarity. *Bulletin of the Psychonomic Society* 12, 433–436.

Straube, Sirko, & Fahle, Manfred (2011). Visual detection and identification are not the same: Evidence from psychophysics and fMRI. *Brain and Cognition*, *75*(1), 29–38.

Tagliabue, C. F., Mazzi, C., Bagattini, C., and Savazzi, S. (2016). Early local activity in temporal areas reflects graded content of visual perception. *Frontiers in Psychology,* 7(APR):1–10.

Titchener, E. B. (1905). *Experimental Psychology: a manual of laboratory practice. Vol. II Quantitative Experiments, Pt. I Student's Manual.* The Macmillan Company, New York

Thomas, James P. (1983). Underlying Psychometric Function for Detecting Gratings and Identifying Spatial Frequency. *Journal of the Optical Societyof America*, *73*(6), 751–758.

Thomas, James P. (1985). Detection and identification: how are they related? *Journal of the Optical Society of America*, *2*(9), 1457.

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

Weiskrantz, Lawrence (2009) *Blindsight: A Case Study Spanning 35 Years and New Developments*. Oxford: Oxford University Press.

Wierzchoń, Michal, Paulewicz, Borysław, Asanowicz, Dariusz, Timmermans, Bert, & Cleeremans, Axel (2014). Different subjective awareness measures demonstrate the influence of visual identification on perceptual awareness ratings. *Consciousness and Cognition*, 27(1), 109–120.

Windey, B., Vermeiren, A., Atas, A., & Cleeremans, A. (2014). The graded and dichotomous nature of visual awareness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1641), 20130282–20130282.

Xu, Yaoda (2010). The neural fate of task-irrelevant features in object-based processing. *Journal of Neuroscience*, 30,14020–14028.

Yeshurun, Yaffa, Carrasco, Marisa, & Maloney, Laurence T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research*, *48*(17), 1837–1851.

Yoshida, M., & Isa, T. (2015). Signal detection analysis of blindsight in monkeys. *Scientific Reports*, *5*, 1–11.

Zadbood, Asieh, Lee, Sang-Hun, & Blake, Randolph (2011). Stimulus Fractionation by Interocular Suppression. *Frontiers in Human Neuroscience*, 5(November), 1–9.

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