The role of conscious awareness in perceptual learning

Adelina-Mihaela Halchin

Thesis submitted to Cardiff University for the degree of Doctor of Philosophy (PhD) in Psychology

December 2023

Contents

Summary

Adaptively learning from encountered visual information is a fundamental skill that can translate into changes at both the behavioural level and at the subjective, conscious experience level. However, the conditions needed for such visual perceptual learning (VPL) to take place are not well understood. It has been proposed that VPL can occur even when the visual information driving the learning is removed from awareness – a finding with implications about the scope of unconscious processing and the function(s) of consciousness. This thesis seeks, first, to re-evaluate the claim that VPL can occur from unconscious information. In Chapter 1, I provide an overview of the literature on VPL and the challenges of studying effects from unconscious visual information. Chapters 2 and 3 detail novel experimental work attempting to drive two different kinds of VPL with unconscious information in different paradigms: two-tone image disambiguation (Chapter 2) and contrast discrimination learning (Chapter 3). The learning conditions in both chapters are contrasted with carefully selected control conditions. In both chapters, there was Bayesian evidence that learning occurred under some experimental circumstances, but crucially, there was no differential advantage of the learning compared to the control conditions. This pattern suggests that the observed learning effects could not be attributed exclusively to the training on unconscious stimuli. Each chapter includes considerations about measurements of awareness and/or learning. Chapter 4 specifically turns towards the issue of measuring (lack of) consciousness, to explore how changes in a widely used subjective measurement, the Perceptual Awareness Scale, relate to changes in task performance. Collating and re-analysing data from Chapters 2 and 3, alongside datasets from 11 published articles, the results highlighted substantial heterogeneity across studies in the relationship between PAS and task performance. Altogether, as discussed in Chapter 5, these results challenge previous conclusions of VPL from unconscious information, and provide rich explorations of whether and how different experimental design choices impact conclusions about awareness and learning.

List of figures

List of tables

List of abbreviations

AB = attentional blink

AFC = alternative forced choice

ANOVA = analysis of variance

BF = Bayes factor

BF_{alt} = Bayes factor for the alternative hypothesis

BF_{null} = Bayes factor for the null hypothesis

C = conscious condition

CFS = continuous flash suppression

CR = confidence ratings

DRDs = dynamic random-dot displays

EC = exclusion criterion

FOW = feeling-of-warmth scale

IFC = interval forced choice

ISI = interstimulus interval

MCQ = multiple choice question

MC = mostly conscious condition

MoC = method of constant stimuli

MoL = method of limits

MU =mostly unconscious condition

PAS = Perceptual Awareness Scale

PDW = post-decision wagering

PF = psychometric function

RSVP = rapid series visual presentation

SD = standard deviation

SEM = standard error of the mean

SOA = stimulus onset asynchrony

SNR = stimulus-to-noise ratio

TIPL = task-irrelevant perceptual learning

VPL = visual perceptual learning

U = unconscious condition

Acknowledgements

Between the challenges and isolation of the pandemic, dealing with anxiety, and a few rounds of burnout, the course of this degree has been far from smooth. Yet, I have been fortunate enough to not have to know how it would be to do a PhD without a strong support system. It really takes a village; here is mine, with my wholehearted thanks.

To my supervisor Aline Bompas, who has been such a reliable and grounding presence, and whose kind straightforwardness I hope I will be able to emulate one day. Thank you for always challenging me with new perspectives and for setting the standard for rigour in thinking – it taught me so much about how to be a better scientist, and I certainly will think back to this experience throughout my career.

To my supervisor Christoph Teufel, whose lectures during my Bachelor's first showed me that perception is pretty cool, and whose support at the beginning of my research training encouraged me to pursue a PhD in Cardiff. I am glad I did, and I am thankful for your guidance throughout.

To my late supervisor Prof. Bill Macken, whose effervescent passion for science guided me in the first few months of the degree, and showed me the magic and fun of 'not knowing…yet' – a perspective that I have tried to keep in the back of my mind ever since.

To Roberto, who had a daily front-row seat to the highs and lows of this journey, and who showed up, steadfastly and with seemingly endless patience every time I needed it, with encouraging words and a listening ear, (a lot of) coffee, and so much more. I am beyond grateful for you.

To the fantastic people that this degree has brought around me, by sheer circumstance and tremendous luck: Hellen, Teodor, Francesco, Abi, Aminette. It has been such a privilege to navigate the last few years alongside you, and I am looking forward to keeping in touch across many kilometres, international borders, and career changes.

And finally, to the brilliant early-career consciousness researchers that I have met and worked with along the way, who gave me a much-needed reminder of the power of community and exchanging ideas about this fascinating puzzle that is consciousness. I am confident that the topic, with its multifaceted challenges, is in good hands.

Data collection for Chapter 2 has been conducted with the help of Dea Bajrami. Validation of the free-naming responses was conducted with the help of Philip Schmid as Rater 2.

Data collection for Chapter 3 has been conducted with the help of Aline Bompas.

Statement of impact

Chapter 2 (introduction, experimental design and methods, and pilot data) was submitted as a Registered Report, and it received Stage 1 In-Principle Acceptance from Peer Committee in Registered Reports. The content of Chapter 2 contains deviations from the Stage 1 approval.

Halchin, A.-M., Teufel, C., & Bompas, A. (2023). *Can one-shot learning be elicited from unconscious information?* https://doi.org/10.17605/OSF.IO/JUCKG

1 Chapter 1

1.1 General introduction

The human visual system has the remarkable ability to tune itself to the information it is exposed to, with such changes often being maintained for a long time. These long-term changes in subjective experience and performance that occur adaptively based on demands from the environment (Gibson, 1969; Sagi, 2011) have been defined as visual perceptual learning (VPL). While VPL can be and has been intuitively linked to the development of visual skills in childhood (Adolph & Kretch, 2015; Chapters 16-19 in Gibson, 1969; Goldstone, 1998), there are many real-life examples of adults learning new highly-specific skills with training: assessing woven fabrics from tactile information (Civille & Dus, 1990), evaluating chest x-rays for disease (E. M. Kok et al., 2012), detecting dangerous items in airport security scans (McCarley et al., 2004; Sowden et al., 2000). More than that, learning can be induced even in fundamental visual skills in the laboratory through extensive practice (e.g., Bao et al., 2010; Fine & Jacobs, 2002; Furmanski et al., 2004; Furmanski & Engel, 2000; Schwiedrzik et al., 2009, 2011; Sowden et al., 2002; Watanabe & Sasaki, 2015), recognized already more than a century ago as "*the progressive mastery of certain simple percepts*" (Judd & Cowling, 1907, p. 349).

What sets VPL apart from other types of learning, like procedural learning (of motion sequences, like driving or walking), associative learning (between a stimulus and a response), or declarative learning (of factual information, like important dates in history) is that VPL is inextricably intertwined with conscious experience. This has been recognized early on, for example by Judd and Cowling (1907) who proposed that*:*

> *whenever a change appears in conscious experience as a result of practice, the elements of the experience are sure to undergo a rearrangement of such a character that they will be more easily discovered than a relatively static experience which is undergoing no marked development.* (p. 350)

Later conceptualizations built upon the same ideas that place changes in awareness at the very core of VPL, with Gibson (1969) describing that (V)PL *"has a phenomenal aspect, the awareness of events presently occurring in the organism's immediate surroundings. It has also a responsive aspect; it entails discriminative, selective responses to the stimuli in the*

immediate environment" (p. 3). In other words, to develop VPL it is not sufficient to memorize a category label, or that a certain sensory input is labelled in a specific way and to respond in a specific way associated with the input; the input must also "*feel like something else"* (Nagel, 1974) after training, as opposed to before.

Anecdotally, this statement is non-controversial and readily supported by everyday examples in both vision and other sensory modalities: for instance, those that do not play the piano look at the keyboard and only see black and white pieces, while piano players see the notes and are able to differentiate between keys only by looking at them. Similarly, those who drink coffee would be able to tell the difference between their favourite brand and a different one, while to a non-coffee drinker both cups would likely taste all the same. Scientifically though, the relationship between PL and awareness has not been explored in depth. Most of the literature focused on the ways in which VPL can exert influence over conscious experience, in a flexible manner tuned to task-specific requirements. Four types of effects were proposed (Goldstone, 1998): differentiation, unitization, attentional weighting, and stimulus imprinting. Below I briefly present each type.

1.2 Differential effects of perceptual learning on conscious experience

Differentiation effects refer to the gained ability to discriminate between stimuli, or to identify new dimensions of the stimulus, that were previously indiscriminable or unidentifiable – with effects spanning from simple stimuli differing in low-level visual properties, such as luminance contrast or orientation, to features or entire complex object categories such as faces. Consequently, classic psychophysical studies on VPL such as detecting which of two consecutive intervals contained a grating of a specific orientation and spatial frequency and which was blank (e.g., Fiorentini & Berardi, 1980), or sharpening of vernier acuity (a decrease of the smallest identifiable degree of misalignment between two lines; McKee & Westheimer, 1978; Poggio et al., 1992), would fall under this category (for reviews, see Goldstone, 1998; Lu & Dosher, 2022). Similarly, learning to break down an initially homogeneous category of complex stimuli (e.g., mushrooms) into functionally distinct subcategories (e.g., poisonous vs. edible) can also be viewed as differentiation.

Conversely, VPL can also lead to *unitization*, or the creation of 'functional units' (Goldstone, 1998) from individual parts previously analysed separately. Holistic processing of familiar faces is a prime example of unitization (Markant & Scott, 2018): a strong relationship was found between familiarity with a face and processing the face holistically over analysing each part separately (DeGutis et al., 2014). Unitization can also occur for other categories of stimuli, like novel complex objects (Liang et al., 2020). The emergence of complex visual categories with well-defined meanings is complemented (or perhaps, supported) by experience-induced unitization in more basic tasks such as learning to detect edges and contours or figure-ground segregation – which was found both across development and in adults (Bhatt & Quinn, 2011; Quinn & Bhatt, 2005, 2006; for a review, see Wagemans et al., 2012).

Attentional weighting refers to learning how to allocate attention, based on previous experience with what aspects of the input predicted correct categorization (Goldstone, 1998; Goldstone & Byrge, 2015; Ransom, 2020). Learning to redirect attention can manifest in sampling different locations after learning compared to before; there is evidence that experts with a category exhibit drastically different patterns of eye moments when looking at an object in their category of expertise, compared to novices or categories outside their expertise (e.g., Brams et al., 2019; Gegenfurtner et al., 2011; Heisz & Shore, 2008). It is worth noting though that attentional weighting is closely linked to unitization and differentiation (Prettyman, 2019): the re-parsing of the input in larger (unitization) or smaller (differentiation) units must co-occur with how attention is allocated to these units, otherwise such re-parsing is not beneficial. However, whether they develop simultaneously, sequentially, or independent from one another is not yet understood.

Finally, *stimulus imprinting* refers to the development of new specific detectors for whole or parts of stimuli (Goldstone, 1998), and has been likened to unsupervised learning, in the sense that categories or relevant patterns can be identified and stored without the need for explicit labels (Goldstone & Byrge, 2015). This type of learning is readily exemplified by two-tone, or Mooney, images (hereafter referred to as two-tone images), like the well-known Dalmatian or the image in Figure 1 in Chapter 2. These images are obtained artificially from manipulating images of real-life scenes or objects (Figure 2), and look like meaningless black-and-white patches to untrained observers who did not see the original, unedited template

image (hereafter referred to as template or greyscale, although these images need not be in greyscale). After exposure to the corresponding template, two-tone images can be perceived as depicting the content of the corresponding template – an effect hereafter called disambiguation. Compared to the other types of VPL discussed, stimulus imprinting can require substantially less practice than learning in other tasks – in the case of two-tones, it can be induced by as little as one exposure to the template image, or in the case of highly-trained categories like faces it can reliably occur spontaneously (Schwiedrzik et al., 2018). Nevertheless, this effect still satisfies the criteria of VPL: disambiguation is substantially facilitated by exposure to the templates (Ludmer et al., 2011) - but not entirely dependent on it, as it can occur spontaneously in untrained images or through exposure to image content labels (e.g., Samaha et al., 2018); it can still be observed in the same images long after the initial exposure (Ludmer et al., 2011); and it alters phenomenology rather than solely driving a stimulus-response association.

1.3 VPL without stimulus awareness?

Fewer studies though have investigated the opposite relationship, namely how awareness might contribute to and whether it is necessary for VPL. As discussed in Chapter 3 (section 3.1), addressing this question is important towards better understanding what cognitive processes might require awareness. Consequently, it might also be relevant towards different theories of consciousness. An overview of different theories of consciousness relevant to VPL, along with a discussion of whether unconscious VPL would be consistent with different theoretical predictions, is covered in Chapter 3 (sections 3.1 and 3.5). A full review of different theories of consciousness is outside the scope of this chapter, but can be found in other sources (e.g., Seth & Bayne, 2022).

This question is different from whether unconscious stimuli can influence behaviour in a masked priming context. VPL effects are thought to be different than masked priming, with a key difference being that with priming, the influence on the subsequent stimuli has been shown to be transient, while VPL is conceptualized as more sustained, longer-term learning. Lin and colleagues demonstrated this distinction in a learning paradigm which combined ‘easy’ trials (presented for over 200ms and allowing higher accuracy, around 8% of trials)

and ‘hard’ trials (presented for 16.7ms and only allowing lower accuracy). Both types of trials had the same structure, namely a square or a diamond followed by a mask. Indeed, in the blocks which contained an equal number of easy and hard trials, there was an increase in accuracy in the hard trials. This advantage started to decay immediately after the easy trials were removed (i.e., in subsequent blocks with only hard trials) and performance returned to baseline after around 10-15 trials. A general improvement throughout the task was also found. In comparison, a control condition which trained only with hard trials also showed a slower, sustained improvement throughout the task, albeit smaller than when easy trials were present. The authors further suggested that, since the learning and priming rates were not found to correlate, the two processes might rely on different underlying plasticity mechanisms (Lin et al., 2017).

The same distinction between masked priming and learning can be applied to stimulus imprinting as well, where comparably fewer trials are needed. For example, in a study using two-tone images, Chang and colleagues (2016) displayed blocks of multiple greyscale images interleaved with blocks of multiple two-tone images (similar to the experiments I conducted in Chapter 2) and found no advantage on performance in two-tone images when the corresponding greyscales were displayed as temporally close as the paradigm allowed (over 4.5s). This finding suggests that on that multi-second timescale, any possible response priming from the masked greyscale had decayed. It can be further argued that response priming does not explain two-tone disambiguation effects outside of this context either, when the greyscale images are visible and presented in a way clearly associated with the two-tones. In a blocked design with 10 two-tones in the pre-post disambiguation stages separated by a long period (60+ seconds) of disambiguation (Teufel et al., 2015), participants’ ability to judge if a two-tone contained a person or not increased.

In any case, the question of ‘VPL without awareness’ can be viewed from multiple angles, based on what stimulus dimension(s) participants are not supposed to be aware of:

1. full awareness of the stimulus and all its low level properties, but not knowing the correct label or the relationship between stimuli; this includes the seminal work from Seitz, Watanabe, and colleagues on task-irrelevant VPL (TIPL) in dynamic random-dot displays (DRDs, Watanabe et al., 2001), because the long stimulus presentation times (500ms) meant that participants were aware of the DRDs and that the dots were

moving, just that the percent of coherent motion was below the threshold for identifying the motion as coherent.

2. no awareness of the target dimension but, at least in theory, possible awareness of some other dimensions of the stimuli, like whether something was presented or not (Axelrod & Rees, 2014; Chang et al., 2016; Nishina et al., 2007; Schwiedrzik et al., 2009, 2011; Seitz et al., 2005);
3. no awareness of the entire stimulus, as in not being able to detect that a stimulus was presented at all (Experiment 2 in Seitz et al., 2009).

Because the conceptualization of '(un)awareness' that I adopt in the thesis involves attempts to remove at least some low-level properties from awareness, the subsequent discussion will target literature from the 2nd and 3rd categories above. It will also only target experiments where during training, participants were exposed exclusively to unconscious trials, bar for attention checks/catch trials. This is because it would otherwise be difficult to establish if any learning effect was driven by the conscious or unconscious trials.

The findings in most studies which meet these criteria indeed seem to support that learning can occur when the target dimension or the whole stimulus is not perceived, in a variety of tasks. Seitz and colleagues (2009, Experiment 2) compared orientation discrimination performance before and after participants trained for 20 days on oriented gratings rendered unconscious by continuous flash suppression (CFS). In CFS, the stimuli to be rendered unconscious are presented to only one eye, while the input to the other eye is a strong mask usually comprised from complex geometric patterns ('Mondrians', see Figure 2 in Seitz et al., 2009, for a visualization). Due to the mismatch between inputs, what is usually consciously experienced is the mask, rather than the stimuli – hence why CFS has been described as related to binocular rivalry (Tsuchiya & Koch, 2005). The gratings were presented in noise, with the signal-to-noise ratio (SNR) either varying (pre-post sessions) or fixed at a specific value (training sessions). The training itself consisted of passively looking at a screen while gratings-in-CFS at the chosen SNR appeared occasionally, either with a specific orientation (trained gratings, 160 trials per training session) or a different orientation (control gratings, 160 trials). All trials containing a grating, either trained or control, were presented to the same eye (trained eye) – a second set of trials involved presenting a blank screen to the other eye (untrained eye). The presentation of the unconscious trained gratings was paired

during training with the delivery of a drop of water, which was rewarding to the participants because they were asked to not eat or drink for 5h before each session. An increase in orientation discrimination accuracy from before to after training was found, only for orientations paired with a reward during training and only for the trained eye. Interestingly, the learning manifested only for the SNR yielding the lowest accuracies before training, and not the SNR used in training, which was already at ceiling before training – thus suggesting that orientation processing as a whole improved in the trained eye, rather than processing of the specific trained stimulus. In a different study on task-irrelevant perceptual learning (TIPL), participants were repeatedly exposed to gratings-in-noise stimuli at different locations in the periphery, while doing an attention-demanding rapid series visual presentation (RSVP) task (Nishina et al., 2007). The RSVP task required participants to pay attention to a continuous stream of characters (letters, and two numbers), and indicate at the end of the trial which numbers were presented. The stream was presented in the periphery as well, not overlapped with the gratings, and participants were required to maintain fixation in the centre of the screen. The contrast of the gratings was chosen to yield chance orientation discrimination pre-training, on average across participants. Performance improved pre-post training for the orientation paired with a target, and there was also an effect of proximity with the target. The findings were interpreted as showing that repeated exposure to unattended, initially indiscriminable gratings can lead to improvements when presented simultaneously with task-relevant stimuli. Another similar TIPL paradigm also found a reduction in the thresholds of psychometric functions for detecting gratings of different contrasts from pre- to post-training on a subthreshold grating, which was modulated by whether the gratings were paired during training with a high-reward or low-reward target (Pascucci et al., 2015). Nevertheless, I will discuss in Chapter 5 (section 5.2) a different perspective on these results (in the context of the specific methodological choices of the studies and of findings from the present thesis), that contests the interpretation that they were exclusively driven by the unconscious stimuli. Other findings, discussed in detail in Chapters 2 and 3, also argued support for VPL from masked visual stimuli, including Mooney image disambiguation from masked templates (Chang et al., 2016), and increases in discrimination sensitivity and PAS from initially-indiscriminable shapes (Schwiedrzik et al., 2009, 2011).

Nevertheless, some exceptions were reported. In an attentional blink (AB) paradigm (Seitz et al., 2005), DRDs with 5% motion coherence were presented peripherally and were task-irrelevant, while participants completed a RSVP task with letters and numbers. The setup of the AB task manipulates the stimulus onset asynchrony (SOA) between two targets (T1 and T2) in the RSVP stream, so that when T2 appears after a certain SOA after an identified T1, participants are less likely to consciously experience T2. For comparison, a Non-AB condition was introduced, where the SOA between T1 and T2 was sufficiently large to enable conscious experience of both targets. Seitz and colleagues (2005) found that in the Non-AB condition where participants were aware of both the DRDs and the T2 target but only attended the target, motion coherence detection in the trained direction improved across 10 training sessions. In the AB condition, where both the DRDs and T2 targets were reportedly unconscious, there was no improvement in motion coherence detection. One caveat however is that both awareness and attention were absent in the experimental AB condition, making it impossible to distinguish between needing awareness of or attention to the stimuli. Similar results of no learning were obtained in a study on holistic face processing (Axelrod & Rees, 2014, Experiment 2). Participants saw two composite face stimuli in each trial, where the eye region was always visible, while the visibility of the rest of the face was manipulated using CFS. The eyes could be aligned with the face thus enabling holistic processing, or slightly shifted thus reducing holistic processing. If holistic processing was present, it would be expected that judgments about the eyes area alone would be slower and less accurate because of the influence of the rest of the face. The training consisted of repeated exposure to the same composite stimuli where the visible eyes were aligned with the invisible faces and both could be the same or different, while participants judged if the eyes alone were the same or different. If the unconscious training strengthened holistic processing, then an effect should be observed on the aligned stimuli only. Participants' accuracy in discriminating if the two faces under CFS were the same or different (while the eyes stayed the same) was not significantly different from 0, neither before nor after training – suggesting that participants were not initially aware of the faces and the training did not change this. Moreover, while some general training effect was observed, there was no significant difference between the aligned and shifted conditions – suggesting that the improvement could not have come from the unconscious faces. In other studies, the learning effect was influenced by the presence of 'easy' trials (i.e., at stimulus parameters

likely to be accurate) and/or feedback (Lin et al., 2017; Liu et al., 2012). For example, Liu and colleagues (2012) showed that the contrast necessary for reaching 65% accuracy did not decrease with training (i.e., no VPL occurred) when the training included no feedback and consisted exclusively of trials at contrasts yielding 65% accuracy in a staircase – but it did decrease when a substantial proportion (i.e., half) of the trials were at 'easy' contrasts (yielding 85% accuracy). However, it is questionable whether the stimuli were unconscious in the beginning of the training, since neither study had the goal of starting the training with unconscious stimuli.

1.4 Methodological considerations in studying unconscious learning

One considerable complication in the investigation of *any* effect of unconscious information on cognitive processes or decision making (or, in the case of this thesis, VPL) is that it requires demonstrating that in a specific condition or set of trials, participants had no conscious awareness of the target stimulus or stimulus feature. Attempting to do this though is far from straightforward, and is riddled with methodological, theoretical and measurement challenges (e.g., Balsdon & Azzopardi, 2015; Meyen et al., 2022; Michel, 2022; Reingold & Merikle, 1988; Sandberg et al., 2010; Sandberg & Overgaard, 2015; Skóra et al., 2021; Szczepanowski et al., 2013; Timmermans & Cleeremans, 2015; Vadillo et al., 2016; Wierzchoń et al., 2019; Zher-Wen & Yu, 2023). At prima facie, one issue might seem to stem from the plurality of measurements in the field, noted already 35 years ago by Reingold & Merikle (1988) in relation to different types of objective task performance (i.e., tasks or measurements where there is a correct answer that the researcher has access to, such as whether a stimulus was presented or not). With an extensive list of choices, it is plausible that each measurement might tap into slightly different aspects of consciousness, and their conclusions might not be compatible.

To assess if participants consciously experienced a stimulus, a non-exhaustive list of possible approaches includes asking participants:

- if they saw the stimulus or not in a given trial (seen/not seen paradigms, e.g., Overgaard et al., 2006; Sidis, 1898),

- how clearly they experienced the (whole, or specific features of) stimuli, on a 3 or 4-point scale between 'No experience' and 'A clear experience' (the PAS introduced by Ramsøy & Overgaard, 2004),
- to pick the correct answer about a stimulus feature from an array of options (discrimination or identification; e.g., Schwiedrzik et al., 2009, 2011),
- to respond in which part of the trial the stimulus was more visible (interval forced-choice or IFC detection, e.g., preprint from Amerio et al., 2023; Peters & Lau, 2015);
- to indicate how much confidence they had (confidence ratings, CR, e.g., Balsdon & Azzopardi, 2015; Lau & Passingham, 2006), or
- how much money they would bet (post-decision wagering, PDW; e.g., Persaud et al., 2007), or
- how warm they feel (feeling-of-warmth scale, FoW; e.g., Wierzchoń et al., 2012, 2014) towards their answers on other questions about the stimulus.

A full review of the pitfalls and advantages of each measure is outside the scope of this discussion. It is worth noting though that most studies investigating unconscious effects rely on a dissociation between an index of awareness such as the ones mentioned above (which should show no awareness) and the index of performance hypothesised to be influenced by the unconscious stimuli (Merikle et al., 2001; Reingold & Merikle, 1988). This approach is somewhat different than the unconscious learning studies discussed in the previous section (1.3) and in Chapter 3, which have a pre-post design; in Schwiedrzik and colleagues' studies (2009; 2011) the index of awareness was the same as the index of learning (d-prime), while in the others (e.g., Chang et al., 2016; Nishina et al., 2007; Pascucci et al., 2015; Seitz et al., 2009) the index of learning was the pre-post training difference and awareness was assessed or inferred separately. Nevertheless, I will summarize what each method of measuring awareness mentioned above involves, and some important issues to consider.

One simple but controversial approach to measuring consciousness of a stimulus is to directly ask participants to introspect and answer if they saw the stimulus or not. If they claim they did not, but their performance on judging some dimension of the stimulus is above chance, then the interpretation would be that this judgment may not require consciousness. Sidis (1898) reported such a result in three experiments ("Class C-E") with the task to identify the letter or number presented on 10 cards. Despite sitting sufficiently

far from the cards to claim that the characters were *'mere blurred dots'* (p. 169), participants correctly identified the characters in 23%-35% of trials (percentages inferred from the reported number of correct answers divided by the number of total trials), thus higher than the chance level (which, at most, was 10% when participants were told the available options in Class E). This finding was interpreted as showing *"a secondary subwaking self that perceives things which the primary waking self is unable to get at"* (p. 171), which is alluding at the modern terminology of 'unconscious perception'. However, one important issue with this approach (as well as other dichotomous subjective measures) is that it assumes that all conscious experiences can be captured within the dichotomy of either experienced or not experienced (also referred to as all-or-nothing, see Chapter 4 for details). If that assumption is incorrect, then the answers (and subsequent conclusions) are contaminated by noise from experiences with intermediary levels of clarity. One measurement comparison indeed found this pattern, although it was not assessed statistically: objective accuracy was higher in trials labelled as 'not seen' than labelled as 'No Experience' on the PAS, and respectively lower in trials labelled as 'seen' than labelled as 'A clear experience' (Overgaard et al., 2006). By comparison, using a measure which allows intermediate answers between 'not seen' and 'seen' would still be able to detect in theory if conscious contents were all-or-nothing, because answers would cluster around the range ends. Another related limitation, shared with other subjective measures like the PAS or CR, is the criterion problem (e.g., Michel, 2022; Timmermans & Cleeremans, 2015) – namely that participants might not base their subjective answers on the same information, or on the information the experimenter assumes they will. In other words, participants failing to report a stimulus as seen might not necessarily mean they did not see it, but that e.g., they adopted a very conservative criterion, or they preferred pressing a specific button over another etc. Both limitations thus make it difficult to pinpoint what seen/not seen questions capture.

The PAS is discussed in detail throughout Chapter 4 (in particular in sections 4.1 and 4.4). To provide an overview, it was proposed as a direct measure of the clarity of experience (Ramsøy & Overgaard, 2004), thus having a seemingly clear link with the concept of interest, namely consciousness. Its graded structure (usually 3 or 4-point range varying between "No experience" and "A clear experience") was argued to be more appropriate for capturing the nuances of conscious experience than binary seen/not seen measures, and also more

directly linked to consciousness than CRs (see paragraph on CRs below). Despite these suggested advantages, the use and validity of the PAS has been questioned (e.g., Irvine, 2012; Michel, 2019, 2022) – for example because the original (and some subsequent, e.g., the "Almost clear experience" level in Sandberg & Overgaard, 2015) scale descriptions still mentioned confidence-related concepts like guessing in the scale description (e.g., "No experience" was defined as *"No impression of the stimulus. All answers are seen as mere guesses"*, Ramsøy & Overgaard, 2004, p. 12). As mentioned above, PAS is also affected by the criterion problem.

CR asks participants to make judgments about how confident they were in their answers on an objective task, from guessing (or no confidence) to fully confident (for an overview, see Norman & Price, 2015). CRs have been an extensively used measurement in consciousness research, although how the task is phrased and implemented has been variable, e.g., 4-point scales (Balsdon & Azzopardi, 2015; Sandberg et al., 2010; Wierzchoń et al., 2014), dichotomous guess/know options (Dienes & Seth, 2010a; Lau & Passingham, 2006). Because the judgment focuses on the decision rather than on the perception of the stimulus, CRs are metacognitive ('thinking about cognition') judgments. On the one hand, this focus has been criticized for being too dissimilar from the concept of awareness it purportedly measures: in the paper proposing the PAS, Ramsøy and Overgaard, (2004) argued that to them *"there is little or no validity in the claim of equality between certainty of one's report and the level of conscious awareness of a perceptual process"* (p. 8). On the other hand, not relying on introspection about subjective experience can be seen as advantageous in cases where participants might have low introspection skills (Sandberg et al., 2010).

With PDW, participants are given an initial account of a certain monetary value (real, imaginary, or tokens), and told that on each trial, they must bet a certain amount on their objective answer, amount that they would lose if they were wrong or gain if they were correct – under the assumption or instruction that participants want to maximize gains (Persaud et al., 2007; Wierzchoń et al., 2012). The amount they would bet on an answer is then taken to indicate their awareness – and therefore, if their bets did not maximize gains but their performance was above-chance, then this would signal that they performed the task without awareness. While PDW has been used primarily to assess implicit knowledge in tasks such as artificial grammar learning (e.g., Dienes & Seth, 2010a; Persaud et al., 2007), it

also produced interesting results when stimulus awareness was targeted instead: for example, a patient with cortical blindness in the right hemifield was just as likely to bet low as to bet high following correct answers in a visual discrimination task (Persaud et al., 2007). This finding falls within the classical definition of 'blindsight', which refers to cases where patients with primary visual cortex damage can nevertheless make correct judgments about visual stimuli presented in their 'blind' field, whilst denying they consciously experienced the stimuli (Sanders et al., 1974). Because PDW does not directly require introspection about the stimulus or the decision, it was argued to be a more intuitive measure of awareness, and like CRs, circumvent the issue of some participants having poor introspective skills (Koch & Preuschoff, 2007; Persaud et al., 2007). However, PDW was criticized for being susceptible to risk aversion confounds (Dienes & Seth, 2010a), it was found to have less tight links with task accuracy than CR and PAS (Sandberg et al., 2010), and it was argued to target metacognitive judgements similar to CR rather than awareness (Overgaard & Sandberg, 2012) – thus suggesting that it might tap into different processes.

Finally, the FoW scale also involves a 4-point rating, with participants being asked to judge how warm they felt toward the answer on an objective measure (1 – "cold", 2 – "chilly", 3 – "warm", 4 – "hot"; Wierzchoń et al., 2012). This approach is rooted in the classic use of FOW judgments in intuition judgments, under the interpretation that the warmer the ratings the closer participants thought they were to the solution (Metcalfe, 1986). Wierzchoń and colleagues (2012) argued that FOW was a suitable measure of awareness because its target, the *'experience of accuracy-related feeling'* (Wierzchoń et al., 2012, p. 1144) could arguably capture possible partial awareness of the stimulus or task-relevant knowledge that the participants might not be able to verbally describe. However, FOW has not been extensively used as an awareness measure, nor has its links been explored in detail (at a theoretical level, with conceptualizations of consciousness, and at an empirical level, with objective measurements of consciousness), so it is difficult to interpret it in relation to the wider literature.

Another way to establish if participants were aware of a stimulus is to assess their objective ability to make a stimulus-related judgment. A popular method is to ask participants to select an answer from a series of alternatives, in a discrimination task (e.g., whether a stimulus was a square or a diamond or if a natural scene image contained an animal or not).

Then, one could calculate d-prime (a bias-free measure of sensitivity stemming from signal detection theory) and compare it to 0 (Axelrod & Rees, 2014; Schwiedrzik et al., 2009, 2011) or take accuracy percentages and compare them to chance (Bacon-Macé et al., 2005). The interpretation would be that if participants were conscious of the stimuli, then their performance would be higher than chance or null sensitivity – in other words, it assumes that consciousness of the stimuli is a prerequisite for better-than-chance performance. Because they do not rely on subjective reports by definition and thus avoid the issues discussed above, objective measures could be seen as preferable (Irvine, 2012). However, others (e.g., Lau, 2007) disagree with the theoretical stance that task performance should be interpreted as consciousness, on the grounds that very simple systems (i.e., photodiodes) can have d-prime above 0 while arguably not being conscious of the dimensions they measure (i.e., light). Practical issues with favouring objective methods are discussed elsewhere in the thesis (section 4.1).

Another approach uses a n-IFC paradigm in combination with a subjective measure. In IFC paradigms, each trial consists of n intervals that are identical bar for the key task-related difference(s). For example, in Peters and Lau's (2015) 2IFC paradigm, the first interval showed a masked grating while the second interval only showed the mask. Participants responded if the grating in each interval was oriented to the left or to the right, and either on which orientation judgment they would bet (Experiments 1 and 2), or in which interval the grating was more visible (control experiment; results discussed in Chapter 5, section 5.3). A similar task involved judging if a line was misaligned to the left or the right of a control line, with one interval containing no misalignment (Amerio et al., 2023, preprint); the subjective question asked which misalignment was more visible. Here, the consciousness judgment still relies on introspection, so answers are still subject to the limitation of participants' introspective abilities – however, because one of the two intervals does not provide any task-related information, this approach arguably cannot be criticized for being contaminated by participants' criterion differences (or biases).

Given how different these judgments are, it is unclear to what extent they rely on the same underlying mechanisms or on the same concept of consciousness. For example, Overgaard and Sandberg (2012) argued that measures relying on introspection about conscious experience (such as the PAS) might differ substantially from other metacognitive judgments,

like CR or PDW. In the same vein, Overgaard et al. (2006) also argued that dichotomous seen/not seen answers are fundamentally different than those on a measure with intermediary options (like the PAS). Anticipating the discussion in Chapter 5, it might not be possible to know if they all measure equivalent aspects of consciousness, even if biases and alternative explanations could be accounted for. Nevertheless, one avenue would be to attempt systematic comparisons of different measures, to evaluate first the presence and extent of any possible differences in results and conclusions about awareness.

Some studies systematically manipulated and compared different measurements, primarily subjective measures, in the same paradigm and stimuli (Sandberg et al., 2010; Szczepanowski et al., 2013; Wierzchoń et al., 2014), and provided incipient evidence that the best-performing subjective measures (a detailed discussion of what 'best-performing' might mean is included in Chapter 4) differ with experimental design (such as whether it is presented before or after an objective measure) and stimulus type (emotive vs neutral faces). Other studies found that making more conservative the criterion for 'not seen', either through changing the post-hoc labelling if the stimuli in a trial were consciously experienced (Balsdon & Azzopardi, 2015), or changing participants' answer criterion through instructions (Jannati & Di Lollo, 2012), changed the conclusions about whether it is possible for awareness to differ between stimulus parameters that yield the same task performance (a pattern called 'relative blindsight', Lau & Passingham, 2006). Relative blindsight was found when trials where participants answered that they had low confidence in their shape discrimination answer were rated as 'not seen' (criterion less conservative), but not when these trials were rated as 'seen' (criterion more conservative; Balsdon & Azzopardi, 2015). Similarly, it was found when participants were instructed to only answer 'seen' when they 'actually' saw the stimulus but not when they were instructed to answer 'seen' when they 'thought' they saw the stimulus (Jannati & Di Lollo, 2012). Alternatively, one may not need to ask participants anything, and instead use indirect physiological and brain imaging indices of awareness (no-report paradigms, Tsuchiya et al., 2015).

However, this plurality of approaches is less a cause and more a symptom of an equally vast landscape of differences in how awareness is understood and what characteristics it is believed to have, at a theoretical level. Using any measure, and by extension assessing how well a measure captures the measurand (i.e., quantity to be measured), can only be justified

in the context of specific a-priori assumptions about the measurand (Skóra et al., 2021). To exemplify, a theoretical position that the subjective nature of consciousness can only be accurately reflected in introspection would entail designing experiments where lack of awareness is defined as lack of some subjective dimension of experience (subjective measures). This position would be followed by further stances about what aspect of subjective reports capture experience, such as clarity of experience, blanket seen/not seen judgments, metacognitive confidence. Conversely, another theoretical position is that introspection is unreliable and prone to criterion shifts that cannot be evaluated nor controlled for in statistical analyses (Irvine, 2012). In turn, this view would entail designing experiments where lack of awareness is defined as participants not performing better than chance in making judgments about some dimension of the stimulus where a ground-truth answer exists (objective measures). Here as well the theoretical position is followed by further taking a stance about what the target judgment should be, such as whether it is sufficient to be insensitive to the task-relevant feature (e.g., chance shape discrimination when stimuli are squares or diamonds) or whether there should be no sensitivity to any aspects of the stimulus (i.e., chance detection).

A more fundamental assumption, adopted throughout the thesis and in empirical work using any kind of reports to assess consciousness, is that awareness of a stimulus always entails reportability. Whether this is the case has been contested in a theoretical distinction between phenomenal (P) and access (A) consciousness, which entails that it is possible that not all conscious experience is available for report (P-without-A, Block, 1995), or in other words that phenomenological experience is richer than the contents available for report. This conceptual distinction has implications for different theories of consciousness: for example, Block (1995) argued that the Global Workspace theory (covered in section 3.1) conceptualizes consciousness in terms of access only, while the Recurrent Processing theory (also covered in section 3.1) was argued to target phenomenology (M. A. Cohen & Dennett, 2011). One notable difficulty for establishing if P-without-A is possible is measuring that phenomenological experience did occur during the stimulus presentation in cases where all reports (be them subjective and objective) would indicate that it did not. For this reason, the debate of whether phenomenology is richer than reportability could be seen as intractable because P-without-A is unfalsifiable (M. A. Cohen & Dennett, 2011; Knotts et al., 2019;

Kouider et al., 2010). Indeed, it is not clear whether P-without-A is possible, and whether it is supported by empirical evidence (Amir et al., 2023) or not (review by Overgaard, 2018). For example, Overgaard (2018) reviewed different results claiming to be consistent with the existence of P-without-A, such as Sperling's finding that while participants reported subjectively seeing all letters in a briefly-presented array, they could only recall (or access) very few unless specifically retro-cued (Sperling, 1960). Overgaard further argued that in all cases, there might be plausible alternative explanations that do not rely on P-without-A; for example, in the case of Sperling's finding, that the subjective claim of having seen all the letters might be based on weak or partial representations that become stronger with retro-cueing – although different definitions of 'access' might allow the possibility of P-without-A.

Yet, Amir and colleagues (2023) recently claimed that they showed that such dissociation can be demonstrated empirically. The authors conducted a hearing experiment in which pink noise stimuli were simultaneously played alongside other sounds, with the sounds decreasing in intensity gradually until fully removed. Conditional on participants responding that they did not hear anything when the noise alone was presented, the noise was then also turned off, and they responded again on whether they noticed a change. All trials were followed by a 2AFC discrimination task between noise stimuli. The authors argued that this design provides a comparison between 'A-trials' when participants could experience and report on the noise throughout the trial, 'P-trials' where they could not report on it but could detect a change, and 'no-consciousness trials' where no change was detected. The authors found that P-trials did occur, albeit rarely (around 12% of trials), and that discrimination accuracy was substantially higher than chance; for comparison, accuracy was at chance in the no-consciousness trials. Despite this pattern being interpreted as demonstrating P-without-A, it is still subject to the same considerations discussed above with introspection, since both questions used to delimit P-trials asked participants to introspect (albeit generally about their experience, rather than the stimuli in particular). Moreover, such general reports about experience ("did you notice a change?") might still be seen as requiring access to the stimuli in order to determine that a change was present. More research would therefore be needed to establish if the assumption of awareness entailing reportability is challenged by these findings.

Another view relevant to the assumption of awareness entailing reportability that I take throughout the thesis (and consequently, to the possibility of P-without-A) comes from the Partial Awareness Hypothesis (Kouider et al., 2010). In line with previous criticism of the evidence for P-without-A, Kouider and colleagues proposed that rather than being conscious of more than we can report, there are different levels at which information is represented and accessible. Where such representation is fragmented (i.e., only conscious access to these fragments is possible), the missing fragments are filled in to give the illusion of rich phenomenology. Consequently, under this interpretation, P-without-A is not possible and *"the impression of richness is not basic and primary, but is actually a late construct"* (Kouider et al., 2010, p. 302). However, the authors further speculate that it might be possible to observe inability to verbally report on stimuli that were consciously experienced. Some reasons for this could be not finding the appropriate words to verbalize the experience, or because participants might have had awareness during the stimulus presentation but might have lost the ability to verbalize it by the time they had to. Conversely, tasks that Kouider and colleagues call 'non-verbal' (i.e., detection, discrimination, similarity judgments) should not show the same caveat. While it is unclear whether report scales like the PAS present this limitation, the speculated distinction underlines further the importance of collecting multiple measurements of awareness – approach taken throughout the thesis.

Therefore, given the plurality of definitions, operationalizations, and measurements in the current research landscape, it seems wise to empirically test questions such as whether different measurement and experimental design choices might influence conclusions about what processes or judgments do not require stimulus awareness, or how different measurements might relate to each other.

1.5 Thesis and methodology overview

The first two experimental chapters are focused on evaluating in more depth previous findings that perceptual learning can occur from visual information not consciously experienced. To summarize, the novelty of the presented work lies in the expanded and improved methodology of conceptually re-testing claims about unconscious visual perceptual learning, as well as in large-scale testing claims about a widely used subjective

measure of awareness (the PAS) in relation to objective measures. Therefore, these methodological changes are present in both the experimental design, and the measurement and analysis of consciousness and learning. Keeping methodological considerations central in re-assessing unconscious VPL acknowledges the issues discussed in section 1.4.

Chapter 2 targets a finding claiming that the disambiguation of two-tone images (described in 1.2) can be elicited from template images that participants had no conscious recognition of (Chang et al., 2016). Two experiments were conducted to re-evaluate this claim, using two measurements of accuracy, while addressing important design and analysis limitations of the original study. In both experiments, participants were presented with two-tone images, before and after exposure to natural scenes that were either visually congruent (i.e., the original template) or incongruent (i.e., different images of the same categories) with the two-tones. The likelihood of consciously experiencing the content of the natural scene images was experimentally manipulated, and awareness was measured using both an objective (image identification) and a subjective (PAS) measure. Disambiguation of the two-tone images was also measured using both an objective (the same image identification task) and a subjective (meaningfulness ratings) measure. The goal of collecting multiple measures was to assess the impact that different criteria for categorizing conscious and unconscious events may have on conclusions related to the role of unconscious information on learning – for this purpose, the same analyses of two-tone disambiguation were repeated across different ways to judge 'unconsciousness' of the templates (subjective alone, objective alone, or a combination of these measures and experimental SOA manipulation). Results from Bayesian statistics, which allowed quantifying the evidence for the null hypothesis, found that generally there was evidence against an advantage of exposure to visually congruent templates when the images were not consciously perceived. While one of the two experiments showed evidence of an increase in both subjective and objective markers of two-tone disambiguation, this effect did not rely on the exact mapping between the two-tone and template images.

However, two-tone disambiguation is a special instance of VPL, with fast learning resulting in robust high-level image recognition – so it could be the case that the relationship between learning and awareness in this task might be equally special, compared to typical VPL effects that require extended practice. Therefore, in Chapter 3 I focus on a different kind of VPL

effect, namely contrast discrimination. Previous research (Schwiedrzik et al., 2009, 2011) suggested that simple shapes (diamonds/squares) which participants cannot discriminate above chance before training can nevertheless break into consciousness as a consequence of the training. At the same time, it was suggested that only a low number of trials (100-200) was necessary for this shift from unconscious to conscious (i.e., chance to above-chance discrimination) to occur (Schwiedrzik et al., 2009). I expand upon this finding by comparing psychometric functions (PFs) for contrast detection and discrimination of metacontrast-masked simple stimuli (left or right pointing arrows), as well as subjective clarity on the PAS, before and after a 1000-trial training session (Learning group) or no training (Control group).

PF measurements were conducted in the first session through a combination of the method of limits (MoL) and method of constant stimuli (MoC), as detailed in Appendix 5, and only through MoC in the final session. The MoL technique estimates the threshold based on two types of trial sets: ascending and descending (Ehrenstein & Ehrenstein, 1999). Both sets involve gradual changes, at pre-defined steps, to the visibility of a stimulus: ascending sets begin with an invisible stimulus whose visibility increases until participants report that they saw it, while descending sets begin with a visible stimulus whose visibility decreases until participants report that they did not see it. The threshold estimate is given by the average of all the intensities at which participants' answers changed, pooled across types of trial sets – since systematic differences were observed between thresholds based on ascending and descending trials only (Pollack, 1968). While this method has the advantage of ease, it assumes that the true threshold lies between the estimates from the two different types of trial sets – assumption challenged by the argument that possible differences could be seen as indicating a meaningful perceptual phenomenon rather than as artifacts (Hock & Schöner, 2010). To benefit from this ease of implementation whilst mindful of the issue with how to interpret the threshold estimate output, the MoL was used only as a pre-test for determining a suitable range of contrasts for each participant, which was subsequently used in the MoC procedure.

The MoC procedure, described in detail in section 3.2.3, involved displaying the same number of trials for each contrast level from a specified list, in a randomized manner (Ehrenstein & Ehrenstein, 1999). The MoC approach has the benefit of allowing the analysis of the raw data at each intensity level in addition to PF comparisons – which can be useful

for measurement comparisons (as in Chapter 4), especially when there is high granularity of levels with many trials in each (as in Chapter 3). The MoC approach can also keep participants (especially inexperienced observers) more motivated, since some of the trials are easy. However, they require a large number of trials, some at intensity levels that are either clearly visible or clearly invisible, so not as informative towards the aim of finding the threshold (Ehrenstein & Ehrenstein, 1999) – therefore this approach is likely not as efficient as other methods of deriving PFs. The efficiency can be increased through pre-testing each participant, as Ehrenstein and Ehrenstein (1999) suggest, to determine an individual range of intensities around the threshold. This approach is followed in Chapter 3, by using the MoL as a pre-test.

Other approaches exist, in which there is no fixed number of trials at each intensity level. For example, staircases are an adaptation of the MoL (Ehrenstein & Ehrenstein, 1999), in which the intensity of a trial is influenced by answers in the previous trials. For instance, a simple staircase might present a series of stimuli with descending intensity in a row until a participant changes their answer to 'not seen', after which the direction is swapped to ascending and the next trial has a higher intensity (other variations of the procedure exist, where direction changes are also influenced by how many prior trials were answered with 'not seen'). The threshold estimate is then given by the average of the intensities at which the response changes. While staircases could be more efficient than MoC, they are not appropriate for experiments where the data underlying the PFs is important too, because each intensity will have been presented for a different number of trials.

For the Learning group in Chapter 3, the training contrast was chosen from the pre-training discrimination PF to yield under 60% accuracy (more in-depth justification for this value can be found in section 3.2.3). The planned comparisons focused on detection and discrimination inflection points, defined as the stimulus characteristic (here, contrast level) at which the PF is the most variable (Strasburger, 2001). This point need not be the halfway point of the curve, which can be referred to as threshold (i.e., 75% accuracy in a task where performance can only plausibly range between chance at 50% and ceiling at 100% minus the lapse rate) – hence why I refer to it as 'inflection point' in Chapter 3 analyses. Indeed, the specific function that was fitted (Weibull) has its inflection point at ~81% when the guess rate is fixed at 50% and lapse rate at 1% (Kingdom & Prins, 2016, Chapter 4, Box 4.5). The

findings were that both performance (lower inflection points for both detection and discrimination PFs) and subjective experience (higher mean PAS) increased between Pre and Post sessions, indicating that learning occurred. However, there was Bayesian evidence against any differences in PFs between the Learning and Control groups, hence suggesting that the learning could not have been due to repeated exposure to stimuli that were not reliably discriminated initially.

Conducting the research in these two chapters highlighted considerations about how consciousness is measured, and prompted me to consider how different measures might relate to each other. In Chapter 4, I focus directly on assessing the relationship between objective measures of awareness and the PAS, as well as which experimental design factors, if any, influence this relationship. I collated datasets from 13 studies, either openly available on the Open Science Foundation repository or shared privately by authors, in addition to data from Chapters 2 and 3. The data spanned a variety of tasks, stimuli, masking methods, and research questions, with the common points that all collected trial-by-trial a version of the PAS and an objective performance measure. The goals were two-fold: to verify the degree to which trials rated with PAS1 ("No Experience") yielded, overall, chance performance in the objective task, and evaluate how well changes in PAS ratings mapped onto changes in objective accuracy. These two questions were chosen because they map onto two dimensions identified in the literature as important metrics about the PAS: exhaustiveness and respectively sensitivity, which have to some extent been linked to different facets of validity. The analyses highlighted substantial heterogeneity across experiments in whether subjective ratings of 'No Experience' entailed also chance performance, with most experiments failing to meet this criterion. They also highlighted that while strong links between the PAS and accuracy were present and, at the group level, the evidence was not consistent with the explanation that awareness is all-or-nothing, there was again substantial heterogeneity between participants and across samples. Moreover, I question the soundness of attempting to 'validate' the PAS against objective measures, and discuss the often-implicit assumptions behind this rationale.

Neither empirical chapter restricted participation to only experienced observers – indeed, in Chapter 2, participants were selected partially on the basis that they had not participated before in other experiments using this stimulus set. This approach is markedly different from

other experiments (primarily in psychophysics) that employ a so-called 'small-n design' (Smith & Little, 2018), in which a small number of observers (often 5 or fewer, typically experienced) complete a large number of trials – with each observer providing an internal replication of the results. The authors argued that results from small-n experiments are not inherently less reliable than those from large-n experiments; instead, they are reliable because small-n experiments tend to employ well-defined measurement methods (i.e., PFs) rooted in strong theories that allow precise predictions (as opposed to ordinal predictions such as 'accuracy in condition X will be higher than in condition Y'). While I agree that such elements can be important for increasing the reliability of findings, a few reasons motivated not employing small-n designs. Firstly, Smith and Little (2018) specifically mention that perceptual learning is not typically the main phenomenon of interest in small-n designs, but characterizing the limits of the target system is – the latter aim being more suitable for being studied in already-trained systems (i.e., experienced observers). Secondly and more importantly, consciousness science is arguably not yet in a position to have similarly strong measurements and theories – hence why it might not benefit as much from a small-n approach.

Nevertheless, elements of this approach were introduced where it was possible to do so. In both empirical chapters, participants completed training and quizzing on how to use the introspective PAS measure, to reduce any possible individual differences in the biases (or criterion) of interpreting and using the measure – although it is worth noting that it is difficult to test whether any differences in the use of PAS reflect noise or stable individual differences. Chapter 3 took a psychophysical approach to the question asked by Schwiedrzik and colleagues (2009; 2011), and also included a high number of trials in the MoC task. Compensating a smaller number of participants with more trials per participant was not possible in Chapter 2, given the small, fixed stimulus set. Other aspects, such as frequent attention checks and rigorous exclusion criteria, were also introduced to increase the quality and reliability of the data.

Throughout the thesis, I use Bayesian tests and the resulting Bayes Factors (BFs) to answer the target questions. In the case of the tests used, the strength of the evidence for a specific hypothesis in comparison to another based on the existing data is given by the posterior odds, which is the product of the prior odds (i.e., the odds of each hypothesis prior to

conducting the research) multiplied by the BF (Tendeiro et al., 2024). If the prior odds are assumed to be equal, as assumed throughout the thesis, then BFs are equal to the posterior odds – hence the use of the terminology 'BF' rather than 'posterior odds'. This level of specificity, which p-values do not allow, is essential for cases where support for the null hypothesis is particularly relevant, for example demonstrating that performance was not different between training conditions or from before to after training (or from chance level, as some objective measures of awareness require).

To anticipate the discussion in 2.2.6.1, BFs provide a continuum of evidence with only one value having a predetermined interpretation: BFs of 1 being interpreted as inconclusive (meaning that the data supports as much the null hypothesis as it does the alternative hypothesis). Conventional cutoffs do exist, for example BFs between 3 and 10 taken to indicate 'moderate' (Lee & Wagenmakers, 2014, as cited in Quintana & Williams, 2018), or 'substantial' (Wetzels et al., 2011) evidence. These are acknowledged though to provide a label for convenience only - and ultimately, changes in the label used bear no consequence to the level of evidence. It might also be reasonable to adopt different conceptual cutoffs in different scenarios (Tendeiro et al., 2024); for example, in Chapter 2 where the existence of the learning effect was questioned because of limitations to the original study (section 2.1.1), a higher standard of evidence (BF > 6) was sought than the conventional level (BF > 3) used in Chapters 3 and 4. Despite this interpretation, it is important to note that BFs are not indices of effect sizes in themselves, and therefore they do not inform about how large the differences between groups are (Tendeiro et al., 2024). However, BFs do require defining an a-priori expected distribution of effect sizes (later referred to simply as 'prior') which is factored into the resulting BFs. Therefore, BFs can be interpreted as strength of the evidence in the context of the expected effect size distribution and under the assumption of equal prior odds.

BFs are also conceptually different from p-values, because p-values do not factor in prior knowledge, and the size of the p-value cannot be used to infer strength of evidence: as Dienes (2014) discusses, a high p-value cannot be interpreted as 'very non-significant', just as a low p-value cannot be interpreted as 'very significant'. Therefore, no direct mapping exists between the two values – and although BFs were found to covary with p-values, they were also found to indicate only 'anecdotal' evidence (between 1-3) in 70% of cases where

p-values were significant between 0.01 and 0.05 (Wetzels et al., 2011). This pattern highlights the difficulty of interpreting p-values through a Bayesian lens – and because of this, such attempt is primarily avoided in this thesis. Where a direct comparison is sought with p-values (section 2.7.2), the robustness of the approximation is assessed across different prior effect size distributions.

2 Chapter 2

2.1 Introduction

2.1.1 General overview

Sensory information is noisy and rarely suffices to uniquely specify representations of the world around us. One way to reduce this ambiguity is for the human visual system to combine sensory information with predictions based on prior knowledge of the environment (de Lange et al., 2018; Samaha et al., 2018; Series & Seitz, 2013; Teufel & Fletcher, 2020; Teufel & Nanay, 2017). This notion has been studied in the context of natural scene statistics, where priors are thought to specify long-lasting, global, and context-independent statistical regularities of the environment (Brunswik & Kamiya, 1953; Geisler, 2008). The prior knowledge underlying these perceptual phenomena is thought to be implemented as constraints on bottom-up information processing (Teufel & Fletcher, 2020). Recent evidence, however, also highlights the importance of a different type of prior knowledge, which acts on perceptual processing via top-down modulation. Specifically, several psychophysical (e.g., Christensen et al., 2015; Lupyan, 2017; Neri, 2014; Teufel et al., 2018) and neuroimaging (e.g., González-García et al., 2018; P. Kok et al., 2012) studies in humans indicate how predictions that are based on high-level, context-dependent expectations shape early information-processing in a local, fast, and flexible manner.

Figure 1. Example of a two-tone image.

One example of the influence of context-dependent prior knowledge is provided by two-tone images (1.2). Without relevant prior knowledge, these stimuli are experienced as

collections of meaningless, black-and-white patches (Figure 1). However, once an observer has acquired appropriate prior knowledge, the patches are perceptually organised into meaningful representations of objects. In many studies, prior knowledge is provided in the form of the original undegraded picture, here called template or greyscale image, from which the two-tone image was derived (Figure 2). This effect, although very sudden, can result in a robust, long-lasting change in visual experience that can be observed much later after exposure (Ludmer et al., 2011) – i.e. an example of perceptual learning (Gibson, 1969) through an 'Eureka effect' (e.g., Ahissar & Hochstein, 1997, 2004).

Recently, Chang et al., (2016) suggested that such disambiguation can be achieved even from very brief, masked greyscale images that participants had no conscious awareness of seeing (Chang et al., 2016). To briefly summarize the study by Chang et al. (2016), subjective identification (referred to as "recognition", i.e., perceived ability to discern what is depicted in the image) and objective identification of two-tone images were assessed before and after exposure to greyscale images that were backward masked to disrupt conscious processing. Following each masked greyscale image, participants indicated whether they could identify its content (yes or no answer), with these responses post-hoc used to label the greyscales as conscious/unconscious. As expected, consciously perceived greyscale images led to increased subjective and objective identification rates of two-tone images from Pre- to Post-Exposure. Surprisingly, even when observers reported not being able to identify the greyscale image, their objective performance in identifying two-tone images still improved from Pre- to Post-Exposure, beyond what was observed for two-tone images for which the corresponding greyscale image was never presented (catch two-tones trials). More surprisingly, *subjective* identification of the content did *not* improve, suggesting it was not affected by unconscious priors. These findings were interpreted as evidence that unconscious processing of a greyscale image suffices to build up prior knowledge that can subsequently guide perceptual organisation of the associated two-tone image, leading to improved performance. Additionally, the data suggest different effects of unconscious prior knowledge on objective performance and subjective experience. Given how surprising these findings are, a re-assessment of the conclusions is warranted. Firstly, the conclusions from Chang and colleagues (2016) are evaluated within the context of the recent literature

pertaining to similar concepts. Several limitations in the original study are then addressed, and how the current work intends to re-evaluate this claim.

2.1.2 Learning from unconscious information?

The field of consciousness research is marked by a high heterogeneity of approaches (1.4) and, therefore, drawing conclusions across studies remains a tentative process. However, focusing on the few studies that are most closely related to the target question, recent evidence suggests that the processing of images made unconscious by masking is limited and would be insufficient to support semantic priming or the learning of statistical regularities. For example, Stein et al. (2020) failed to replicate Van den Bussche et al. (2009)'s finding, that forward- and backward-masked drawings of animals and objects primed correct categorization of animal/object words from the same category but different identity (e.g., a masked drawing of a bear facilitated answering 'animal' when presented with the word 'dog'). This replication failure occurred despite using the same stimuli and task as the original authors, with more than double the sample size. Moreover, the authors reported that employing continuous flash suppression (CFS) to render images unconscious did not result in an effect either, despite CFS being thought to allow higher-level processing than sandwich-masking (Breitmeyer, 2015) – casting doubt on the robustness or true effect size of the effect originally reported by Van den Bussche and colleagues. Finally, another similar study showed that when the prime and target had the same identity, backward-masked pictures did not prime basic category (e.g., bear) answers (Koivisto & Rientamo, 2016). Koivisto and Rientamo (2016) did find weak priming for superordinate answers (e.g., 'animal'). However, this small effect may have been driven by a subset of trials that were incorrectly labelled as 'unconscious'. This is not unlikely, since participants had above-chance prime discrimination sensitivity (d-prime) in the superordinate condition (but not in the basic category condition). Altogether, these findings suggest that briefly presented, masked images do not activate sufficient object category information to facilitate the processing of subsequent category-relevant words or pictures.

Other relevant research refers to longer-term learning effects, where an increase in accuracy results from repeated exposure to a type of stimuli over time. This literature typically relies on simpler stimuli such as geometrical shapes or letters, rather than images of natural

scenes or objects. Recent findings from this field also underline that at least some forms of this learning require conscious processing. For example, encoding statistical contingencies about the stimulus location in a masked cueing task requires stimulus awareness (Travers et al., 2018). Likewise, when measuring (un)consciousness of the stimuli with a sensitive task and valid statistical methods (i.e., not inferring evidence for absence of awareness from absence of evidence), awareness of a visual stimulus appears necessary both for encoding the relationship between the stimulus and its rewarding or punitive value (failed replication of Pessiglione et al., 2008 by Skora et al., 2021; Skora & Scott, 2022), and for triggering a preferential response when the relationship was consolidated consciously (Skora & Scott, 2022).

From the above studies, it seems that masking not only prevents conscious access to visual information, but also disrupts high-level processing of this information. In that context, could Chang et al.'s finding be taken to suggest that one-shot learning takes place without high-level processing of the greyscale image? Indeed, one could speculate that the greyscale image leaves some raw visual traces, that would be stored for a few minutes, and retrieved *only* when cued by a consciously presented, visually related image (the two-tone presented post-exposure). Whether this mechanism is plausible or not is, for now, difficult to assess. Furthermore, the mechanisms of backward masking, and its role in consciousness, are still debated and may well be stimulus and task-dependent. Therefore, previous evidence is not sufficient to exclude the possibility that backward-masked natural images can still facilitate future object identification. Conversely, a failure to reproduce this finding would align with the recent literature suggesting that unconscious processing of complex images is limited and does not influence subsequent high-level image subjective identification and categorization. Hence, it seems important to provide further evidence that deals with some of the methodological challenges that hamper interpretation of existing findings.

Figure 2. Example of a greyscale image, corresponding to the two-tone in Figure 1. Photograph sourced from the personal archive of Halchin A.

2.1.3 Limitations of the original study

Because of some of their design choices, the results reported by Chang and colleagues (2016) can be difficult to interpret, as acknowledged by the authors. The original study did not directly manipulate conscious processing of the greyscale images. Rather, in a pilot titration study, they manipulated the duration between onset of the greyscale image and onset of the mask (thereafter referred to as stimulus onset asynchrony, SOA), and asked observers to answer, after each masked greyscale, a yes/no question of whether they felt able to identify the content of the image (Appendix 3). A duration that yielded, on average, equal numbers of 'yes' and 'no' answers in this pilot study was chosen for the main experiment. Subsequently, the same yes/no question was asked during the main experiment, with trials retrospectively divided into "conscious" ("yes" responses) or "unconscious" greyscale images ("no" responses), to test for disambiguation effects. However, this yes/no categorization is likely not sufficiently nuanced, and some participants may well have responded "No" following images they identified the content of but with low confidence.

This leads to two related issues. First, any disambiguation effect could be due to these "partially conscious" trials, and therefore would not constitute clear evidence for perceptual learning from *unconscious* information. Secondly and more importantly, one cannot be certain that any genuine perceptual disambiguation occurred. Indeed, one cannot exclude

that participants may have used these “low confidence - No” contents (alongside all the “yes” ones) to guess answers to the subsequent two-tone images presented at the post-exposure stage, despite the black and white patches remaining meaningless to them. Although participants would use test and catch greyscale images equally when doing so, this strategy would lead to increased accuracy from pre- to post-exposure for their test condition only, not their catch condition. This is because, in the catch condition, the greyscale images were fully unrelated images, and therefore not matching any of the two-tone categories. Therefore, using “partially conscious” contents from catch greyscale images to guess answers at the post-exposure stage would lead to chance performance, while using this same strategy on test greyscale images would lead to above chance performance. In Chang and colleagues’ analyses, this differential boost in accuracy would be indistinguishable from genuine perceptual disambiguation. Participants’ guessing would also explain why identification accuracy, but not subjective recognition, increased pre- to post-exposure, compared to the catch trials.

It therefore cannot be ruled out that the key effect – higher correct identification after unconscious corresponding greyscale images compared to catch trials – might be based exclusively on miscategorised trials. The caveat of conscious information potentially being labelled as unconscious is a fundamental issue for consciousness studies more broadly (Peters & Lau, 2015), and is related (as discussed in Chapter 1) to substantial discrepancies across researchers in their definitions and preferred measurements of consciousness (Eriksson et al., 2020; Francken et al., 2022; Peters et al., 2017), and consequently, the researchers’ conservativeness in their criteria for labelling phenomena as unconscious (Balsdon & Clifford, 2018; Holender, 1986; Vadillo et al., 2016). However, here it is pronounced because the observers’ yes/no answers to the masked greyscale images were not complemented by an objective performance measure.

2.1.4 Overview of the proposed research

Here it was aimed to use a more detailed evaluation of conscious processing to re-assess the notion that prior visual information can be acquired and can operate outside of awareness. Using a conceptually similar method as Chang and colleagues (2016), the current work

tested whether the visual system can organise two-tone images into meaningful percepts after masked greyscale image exposure. Like in the original study, disambiguation was measured using both objective and subjective measures. The key proposed changes are to: (i) experimentally manipulate the extent to which perception of greyscale images is conscious, (ii) use a 4-point subjective greyscale visibility scale, followed by (iii) an objective greyscale identification measure, in order to (iv) provide a more nuanced way to categorise trials into conscious or unconscious, and (v) use a catch condition that attempts to control for guessing. Furthermore, this experiment was repeated using two different objective measures of two-tone identification, and disambiguation (or its absence) was assessed using Bayesian statistics.

2.1.4.1 Manipulating greyscale image visibility

To experimentally manipulate conscious processing, a backward masking paradigm with two different SOAs (e.g., Faivre et al., 2019) was used. The choice of short and long SOAs was aimed at producing weak (e.g., Faivre et al., 2019) and robust awareness respectively and was validated in a pilot experiment (see Pilot 2, 2.4.2).

2.1.4.2 Measuring greyscale image subjective and objective visibility

Instead of a yes/no response, a four-point assessment of participants' subjective experience was collected - the Perceptual Awareness Scale (PAS, Figure 3B in Ramsøy & Overgaard, 2004) - which asked participants to describe the clearness of their perceptual experience from (1) "no experience" to (4) "a clear experience". Objective identification of the masked greyscale images was assessed using two different methods: free-naming (Experiment 1, Figure 3) and a 5-alternative-forced-choice task (Experiment 2).

2.1.4.3 Categorising trials based on greyscale image visibility

Collecting subjective and objective measures, along with direct manipulations of awareness, allowed examining whether conclusions about learning change based on the criteria used for labelling phenomena as conscious or unconscious. Analyses were run on trials classified according to specific combinations of SOA condition, PAS answers, and identification accuracy (Figure 5). The data from two-tone trials, separately for catch and test conditions, was divided into four categories, based on whether the greyscale image was categorised as

fully Unconscious (U), Mostly Unconscious (MU), Mostly Conscious (MC) and fully Conscious (C). These categories and the rationale for them are detailed in the analysis plan and summarised below:

- U: Trials associated with greyscale images that were rated with PAS of 1 ("no experience") and incorrectly identified in the Short SOA condition
- MU: Trials associated with greyscale images that were rated with PAS of 1 or 2 ("a brief glimpse") and incorrectly identified in the Short SOA condition
- MC: Trials associated with greyscale images that were rated with PAS of 2 or higher and correctly identified in the Long SOA condition
- C: Trials associated with greyscale images that were rated as PAS 3 and 4 and correctly identified in the Long SOA condition

At the core of this classification is the assumption that subjective and objective measures of consciousness converge and complement each other. Although blindsight was argued to take place in some circumstances, leading to a PAS rating of 1 but correct identification, the evidence in healthy observers has been challenged (Peters & Lau, 2015; Rajananda et al., 2020), and the pilot data (2.4) described below show that very few trials fall into this category, which would more readily be explained as participant lapses or lucky guesses rather than a substantial pattern to account for.

In addition, analyses were repeated on trials classified using SOA, PAS, and accuracy separately, to contrast conclusions across approaches and allow comparisons with previous research. Tackling this question is important because, in previous literature, all three indices have been used on their own to identify the presence or absence of consciousness – PAS e.g., Koivisto et al. (2013), identification accuracy e.g., Koivisto & Rientamo (2016), SOA e.g., Dehaene et al. (2001) – and whether these approaches lead to consistent conclusions is currently not known.

2.1.4.4. Measuring perceptual disambiguation of two-tone images

Disambiguation was defined as increases in mean accuracy and meaningfulness ratings, from Pre- to Post-Exposure (Figure 4), that are larger in the test condition compared to the catch condition. These comparisons were run within each of the categories described above. The catch condition consisted of two-tone and greyscale image pairs that represent the same

content but come from different images (e.g., two dissimilar images of a peacock). This catch condition did not prevent the guessing strategy, but rather equalise the accuracy benefit between the catch and test conditions. Therefore, a differential effect between the catch and control conditions for both objective and subjective measures would be evidence for an improved ability to *see* the original object amongst the black and white patches following exposure to the greyscale image.

2.2 Experiment 1 – Methods

2.2.1 Participants

91 participants were recruited, in two batches (60, then 31). Only individuals from the general population who were self-declared native English speakers, did not participate in studies using these stimuli before, and without a history of photosensitivity, were recruited from Cardiff University's participant panel, in exchange for payment or course credits. 33 participants were excluded from all analyses for failing one or more attention checks (explained in 2.2.6.3; 11 had accuracies in the visible attention checks under 88%, 27 rated too many visible attention checks with PAS ratings under 3, 6 rated too many blank attention checks with PAS above 1, 6 rated with PAS1 more than half of the Long SOA trials). The final sample was n = 58 (age range 18-22, mean = 19.4, SD = 1.1, 9 males, 46 females, 3 other gender identities). All but 2 of the included participants declared normal or corrected-to-normal vision.

2.2.2 Stimuli

The stimulus set consisted of 23 greyscale template images and their corresponding two-tones images, each representing animals or human beings. An in-depth description of how the stimuli were obtained can be found in Teufel et al. (2015). Duplicates were removed to ensure each content is unique. One additional pair of two-tone and greyscale was introduced to achieve equal block lengths but was removed from all analyses. Each template had an associated 'catch' greyscale, showing a different image of the same content – i.e., if the original template showed a peacock, the catch greyscale would be a different image of a peacock (Figure 3). Another 10 greyscale images were introduced, one at the experiment

instruction stage, and the remaining nine as attention checks (clearly visible images that serve the purpose of excluding inattentive participants from analyses). These 10 novel images, along with the semantically related but visually different catch greyscale images were sourced from personal archive or copyright-free images on pexels.com and edited using Microsoft Office PowerPoint (cropped, and for catch image contrast decreased by 40%). Masks were obtained through phase scrambling the test and catch greyscale images, so each image was associated with its own mask. All images were 3cm x 3cm, sustaining approximately 5x5 degrees of visual angle from the distance of 34cm away from the screen, where participants were instructed to sit (although head position was not restricted). Matching the low-level properties (such as root-mean-square contrast, luminance, edge density, and spatial frequency) between the Catch and Test templates was not attempted, and there was moderate evidence against systematic differences between image pairs, for all properties considered (Appendix 2).

2.2.3 Materials

The experiment was created using PsychoPy3 Builder (Peirce et al., 2019), conducted on the online platform Pavlovia using PsychoJS JavaScript code via Google Chrome, and displayed on iiyama monitors with a diagonal of 21.5-inch, 26.9 cm monitor height, a resolution of 1920x1080, a refresh rate of 60Hz, a contrast set to 80%, and brightness set to its maximum. Two subjective discrete scales were used. First, to assess subjective meaningfulness of the two-tone images, a scale from 1 (not meaningful at all) to 4 (very meaningful) was used, with the question 'How meaningful is this image to you?'. Second, to assess the conscious experience after masked greyscale images, an edited version of the Perceptual Awareness Scale (PAS; Ramsøy & Overgaard, 2004; Sandberg et al., 2010) was used, with the following steps: (1) No experience, (2) A brief glimpse, (3) An almost clear experience, and (4) A clear experience. The instructions related to the scale were slightly modified to focus on the clarity of the experience, rather than confidence in the answer (see Appendix 1). Responses were collected using the keyboard.

2.2.4 Design

Each two-tone was presented twice, once before and once after participants viewed either the corresponding greyscale (test condition), or a catch greyscale (catch condition), as illustrated in Figure 3. The experiment settings constrained the participants to provide an answer for the PAS and meaningfulness ratings, but not the free-naming object identification. The experiment alternated series of two-tone and exposure trials, so that the Pre-Exposure, Exposure and Post-Exposure stages for each image belonged to three consecutive blocks. An illustrative outline of the experiment is included in Appendix 2.

2.2.4.1 Pre-/Post-Exposure stages

Trials in the Pre- and Post-Exposure stages had the same structure (Figure 4A), the only difference being whether they occurred before or after the Exposure stage (Figure 4B). In these trials (Figure 4A), observers viewed two-tone images for 2000ms, allowing conscious perception. After each two-tone image, they were asked to rate its meaningfulness and to complete the object identification task, by typing in a box what they thought the image depicted.

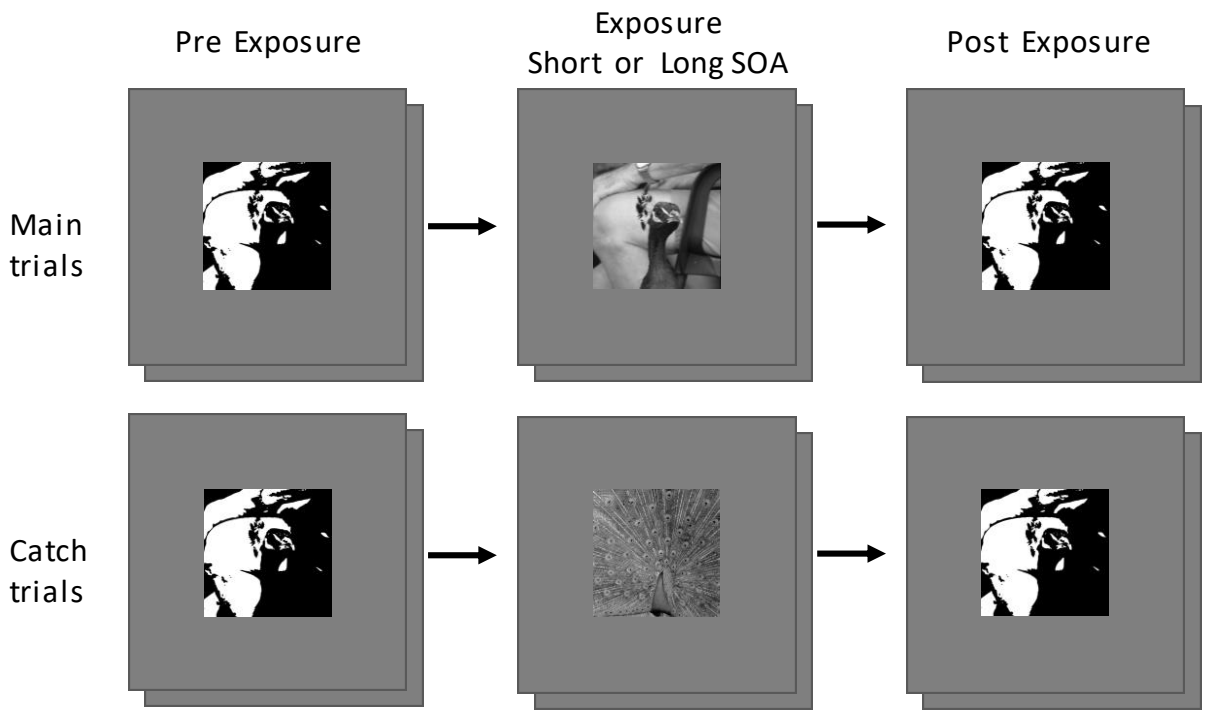


Figure 3. Overview of the experimental design in Experiments 1 and 2. Each square panel represents a trial, as detailed in Figure 4. In the Exposure stage, each greyscale image was

seen in only one of the two masking conditions (Short or Long SOA). For the catch trials in the Exposure stage, a different greyscale image featuring the same object was *presented. Each participant saw a particular two-tone associated with either the corresponding, or the catch greyscale (but never both).*

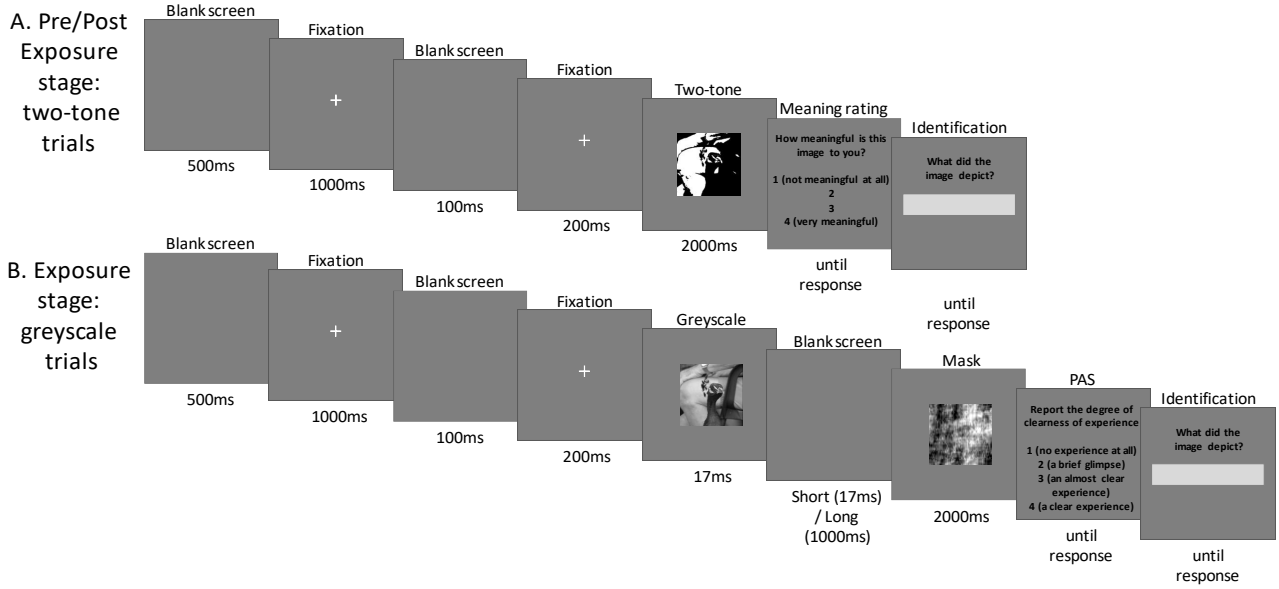


Figure 4. Structure of a trial in the Pre- and Post-Exposure stages (A) and Exposure stage (B) in Experiment 1. A. Each two-tone was followed by a prompt to rate the meaningfulness of the image from 1 (not meaningful at all) to 4 (very meaningful), and to type in a brief description of the content of the image. Each two-tone trial was presented before (Pre) and after (Post) Exposure to a greyscale image. B. In the Exposure stage, greyscale images were *presented for approx. 17ms (16.7), followed by a blank screen and a noise mask for 2000ms. The blank screen had either a short or a long duration. Then, participants rated on the Perceptual Awareness Scale (PAS) how clear they experienced the greyscale image on a scale from 1 (no experience at all) to 4 (a clear experience). Next, they completed the same identification task presented during the two-tone trials.*

2.2.4.2 SOA manipulation

Each participant completed two conditions in the Exposure stage: Short SOA and Long SOA. Both conditions (Figure 4B) started with a 16.7ms display of a greyscale image presented in the centre of the screen. In the Short SOA condition, the image was followed by a 16.7ms

blank screen, and a mask for 2000ms. In the Long SOA condition, the image was followed by a 1000ms blank screen, and a mask for 2000ms. Although this means a longer delay between the image and the response in the Long SOA condition, Pilot 2 (2.4.2) shows that this long SOA consistently led to greater clarity than the short SOA.

Some reports exist that at durations comparable to the Short SOA, participants still showed evidence of having processed the gist of the scene. For example, previous studies reported: high d-prime (above 3) for judging if the content of a 30ms image followed by a mask was congruent or not with that of a subsequent target image (Schyns & Oliva, 1994), ceiling accuracy in category judgments for images presented unmasked for ~26ms (Rousselet et al., 2005), or above-chance accuracy even when masked (~18ms SOA, 6ms image plus 12ms blank, Bacon-Macé et al., 2005). However, one difficulty in comparing the chosen durations with the gist literature is that the visibility or awareness of the images was not assessed in the above-cited studies, so the reported performance could reflect unconscious processing. Moreover, the judgments in these studies (congruency, releasing a button when the picture contained a target scene as in Rousselet et al., 2005, or animals in general as in Bacon-Macé et al., 2005) are different compared to the task in both Experiments 1 and 2. When free-naming descriptions were required (similar to Experiment 1), a 27ms display only yielded vague gist impressions of animate objects (Fei-Fei et al., 2007) – which in the current task would not be sufficient to attract ratings of 'correct'.

2.2.4.3 Template and catch images

Each two-tone image was followed either by its original template image (test condition) or a catch greyscale (control condition). The catch greyscale represented the same object as the two-tone but came from a different image. The catch Exposure trials had the same structure as the test Exposure trials.

2.2.4.4 Attention checks

To ensure only attentive participants who were compliant with the use of the PAS were included in the analyses, two types of attention checks were also included, with stimuli either clearly visible or absent. For the "clearly visible" attention checks, a greyscale image not associated with any two-tone was presented for 1000ms, followed by a blank screen for 1000ms, and a mask for another 2000ms. For the "absent" attention checks, a blank screen was presented instead of the greyscale image. Both types of checks were interleaved with

the rest of the trials in the Exposure stage, and participants completed the PAS and identification questions after each attention check. Each block of greyscale images contained 3 clearly visible checks, and between 1-4 absent attention checks, with a total of 9 clearly visible and 6 absent trials. While both checks ensured that participants could use the PAS accurately, the clearly visible checks had the additional role of attracting correct responses in the identification task, hence why having more of them allowed closer monitoring of the accuracy. Participants were made aware of the presence of attention checks, and the uncertainty on the number of attention checks per block was aimed to keep participants engaged throughout the task by making sure they could not predict how many checks there were in a block.

2.2.4.5 Image and trial randomisation

For each observer, each two-tone image was allocated in a counterbalanced way to one of the 4 possible conditions (2 SOA durations x 2 trial types): Short-Catch, Short-Main, Long-Catch, and Long-Main. Blocks of two-tone or greyscale images alternated (see Appendix 2), and trial order was randomised across and within blocks.

2.2.5 Procedure

The experimental session started with a short training, illustrating the disambiguation effect, and exposure trials with clearly visible greyscale images. Participants were specifically instructed to not leave the free-naming identification box empty even if they did not know the content of the greyscale, and to answer the PAS based on their perception of the contents of the images rather than confidence in the correctness of the answer. Participants were not directly instructed regarding the level of specificity they should categorize the images. However, the instructions stage showed an example image and an example answer ("a deer"), which set the expectation that their answers should be more specific than broad superordinate categories. They had unlimited time during training to study the PAS and the descriptions of each level of awareness and completed an 8-question quiz with feedback to verify their understanding of the scale, before they started the experiment (Appendix 1). Participants were given feedback after each quiz trial and had to repeat the PAS training if they answered incorrectly on more than 1 question. Each experimental session lasted

approximately 25min. At the end of the experiment, participants were asked to type their feedback in a free-response box, and afterwards were debriefed.

2.2.6 Planned analysis protocol and data exclusion

2.2.6.1 Analyses software and statistical choices

Meaningfulness ratings, PAS scores, and accuracy ratings were analysed separately. All data analyses were conducted in RStudio v2021.09.1-372 with R version 4.2.1 (R Core Team, 2021), using the following packages: tidyverse v2.0.0 (Wickham & RStudio, 2023), data.table v1.14.2 (Dowle et al., 2021), here v1.0.1 (Müller & Bryan, 2020), gghalves v0.1.4 (Tiedemann, 2022), R.utils v2.12.0 (Bengtsson, 2022a). All tests used the BayesFactor v0.9.12-4.4 package (Morey et al., 2022; Rouder et al., 2009), to assess the strength of the evidence in favour of the null hypotheses. This is especially important for deciding when participants were not able to use the information presented, which cannot be inferred only from failing to reject the null hypothesis using frequentist statistics (Dienes, 2015, 2016; Vadillo et al., 2016). A BF of 6 (or 1/6), taken to indicate moderate evidence (Lee & Wagenmakers, 2014, as cited in Quintana & Williams, 2018), was chosen (as justified in Chapter 1, section 1.5). It is worth noting though that the strength of the BF evidence is a continuum, without specific thresholds (although BF = 1 signals that the evidence is inconclusive), and that different levels might be considered adequate for different experiments. The BayesFactor package uses a pre-specified Jeffreys uninformative prior on the variance of the population (Rouder et al., 2009), and assumes that the distribution of the standardized effect size follows a Cauchy distribution centred on zero. Following simulations for effect sizes under different prior scales (see Effect size section 2.2.6.2), a wide scale for the prior distribution of effect sizes (r = 1) was chosen, rather than the default r = $\sqrt{2}/2$, because it yielded higher probability of finding evidence for the null if the true effect size was 0 without substantially reducing the probability of finding evidence for the alternative if the effect size was above 0. Moreover, a one-sided prior was chosen, by modifying the prior to be above 0, as only a positive increase (Post Minus Pre, or Test Minus Catch) would be interpreted as support for disambiguation. The simulations indicated that at n= 60, a one-sided prior under r = 1 has minimal impact on false positives and power if the true effect size

was 0, but substantially increases power while reducing the false negative rate for a true effect size of at least 0.30.

2.2.6.2 Effect size and prior distribution estimation

Calculating a Cohen's d value for within-subjects designs can be difficult, as the multiple approaches used in the literature yield substantially different results (Westfall, 2016). For example, using the t-statistic for the comparison between "Grey Not Recognized" and pooled catch trials (page 7, $t = 2.076$, $p = 0.0492$, n= 24), two different effect sizes were obtained: $d_z = 0.42$ (J. Cohen, 1988; Lakens, 2013), and $d_t = 0.6$ through naïve conversion from the t statistic (Dunlap et al., 1996), which is likely inflated due to correlation within the variables. Applying the traditional formula for Cohen's d (which is for between-subjects designs) on points extracted from Figure 4B in Chang et al. (see Appendix 3 for details about their experimental design) yields d = 0.14.

Furthermore, any estimate of an effect size from Chang et al. (2016) would be only marginally useful, given the substantial differences in experimental design and stimulus set. Therefore, to estimate what would be the smallest effect size reliably detectable in the present paradigm, simulations of Bayesian paired t-tests were ran, for different combinations of effect sizes, participant numbers, and prior scales, using the BayesFactor package, along the following parameters:

1. a "medium" scale (default r = √2/2) or a "wide" scale (r = 1);
2. N = 60, 92, 120, 150, 300;
3. Two-sided prior (-r to +r) or one-sided prior (+r), to assess how power and false positives would change if the tests for H1e/f and H2e/f are directional;
4. Possible effect sizes (ES) of 0, 0.14, 0.3, 0.35, 0.42, 0.5, 0.6, 0.707, and 1.

For each combination, 50000 simulations where ran, where an N number of datapoints from a normal distribution with mean = 0 and SD = 1, and a normal distribution with mean = ES and SD = 1 were selected. Bayesian paired t-tests were then ran, with a two-sided or one-sided prior and an R scale, and the percentage of simulations was computed in each combination that resulted in evidence for the Null and the Alternative hypothesis. The code is modified from Lakens (2016).

Under a true effect size of 0 and a one-sided "wide" prior at an n of 60, 87% of simulations resulted in $BF_{null} > 3$, and 71% resulted in $BF_{null} > 6$, with $BF_{alt} > 3$ in less than 1% of cases. Conversely, under a true effect size of 0.5 with the rest of the parameters equal, 63.7% of simulations resulted in $BF_{alt} > 3$, 52.3% in $BF_{alt} > 6$, with 1.4% resulted in $BF_{null} > 6$ (5.49% $BF_{null} > 3$). These parameters deemed the analyses sufficiently powered to detect at least weak evidence for effect sizes of 0.5 and 0.

2.2.6.3 Trial rejection

If a frame drop (e.g., logged duration of 0) was logged in the timestamps from the Pavlovia output for a given greyscale during the Exposure stage, the associated two-tones were removed. Potential intermediate values between 0 and the target stimulus duration of 16.67ms (e.g., 10ms) were attributed to error of logging rather than different stimulus durations and were included. Indeed, while a high percentage of experimental trials had logged durations under half a frame (i.e., under 8.4ms), comparing the distribution of accuracy and PAS answers in these trials found no differences that would indicate dropped frames (i.e., higher proportion of wrong answers or PAS ratings of 1). Under this condition, no trials were removed.

2.2.6.4 Accuracy coding

Each free-naming answer to the identification prompt in the Exposure and two-tone trials was manually validated by two evaluators (A.H., and a Rater 2). Evaluators were blind to whether trials were in the test or catch conditions. Similarly to previous approaches (Samaha et al., 2018), answers were judged as correct if they fully or partially referred to the object, rather than the background of the images. The very brief exposure to the greyscale images may not allow the identification of visual details required to differentiate between visually similar categories (e.g., a hyena and a cheetah). Therefore, answers that also referred to visually similar objects were considered correct. Spelling errors (e.g., 'hayena') and plural forms (e.g., 'hyenas') of correct words were labelled as correct. Answers that were not empty but did not refer to the content of the images, such as 'I don't know', or were too broad and non-descriptive (e.g., "animal") were counted as incorrect. Pairs of images for which participants did not provide any answer (e.g., left the identification box blank) either Pre-, during, or Post-Exposure, were removed. All boxes for the image-absent attention checks, including boxes that were left empty, were by default rated as 'correct' for

convenience, as accuracy in these answers is not analysed and failing these attention checks relies exclusively on the PAS. Inter-rater agreement was very high (96.6% for the post-exclusions final dataset). The remaining cases were discussed between raters until an agreement was reached.

2.2.6.5 Participant rejection

Participants were excluded if they:

1. Failed to reach 88% accuracy (8/9 correct answers) in the clearly-visible attention checks;
2. Answered with PAS 1 or 2 in more than 1/9 clearly-visible attention checks, or PAS above 1 in more than 1/ 6 absent attention checks;
3. Used PAS 3 or 4 in more than half of the trials in the Short SOA condition, or PAS 1 in more than half of the trials in the Long SOA condition.

2.2.6.6 Trial classification based on consciousness of greyscale image

Data from Pre- and Post-Exposure stages were divided based on the SOA manipulation, PAS ratings and identification accuracy during the Exposure stage (Figure 5). In addition to using each criterion alone, they were combined to allow high confidence in the judgments about whether participants had no or partial conscious experience of the stimuli. Four categories were defined, from least (Unconscious/U) to most (Conscious/C) likely that the greyscale images were consciously experienced. On this gradient, Unconscious and Conscious were the most conservative, because they required extreme PAS answers of presence/absence of subjective experience in addition to expected accuracy responses. It is acknowledged that the Unconscious category might have erroneously excluded trials where participants might have guessed correctly by chance despite indicating no awareness otherwise, or conversely for the Conscious category trials where participants might have lapsed despite having awareness of the content. However, although these exclusions might have reduced the number of trials that contributed to the analyses in these two categories, they did not introduce any confounds in the analyses. Including trials with greyscale images rated as either PAS 1 or 2 (A Brief Glimpse) (e.g., Koivisto et al., 2013) made the Mostly Unconscious category less conservative than the Unconscious category because it assumed that even if participants had some awareness to detect that something was presented (and hence report a brief glimpse than no experience), they had no awareness for the purpose of the

identification task (see also Michel, 2022 for a detailed argument in favour of this approach). Similarly, the Mostly Conscious category was less conservative than the Conscious category because of a broader definition of 'conscious' which includes PAS 2. Catch trials were divided by the same criteria.

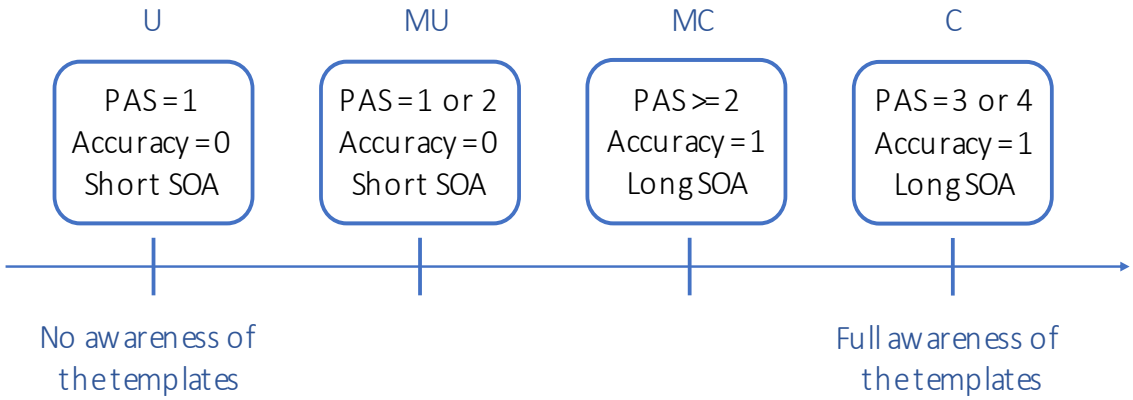


Figure 5. Representation of the different categories (U to C) in which the data were separated, and the criteria that describe each category. The axis at the bottom illustrates expectations about the degree of conscious perception of the images in each category.

2.2.6.7 Planned statistical analyses

Seven hypotheses were tested. H1-H4 followed the format introduced in Table 1. For comparison, the analyses were extended to trials categorised based solely on PAS rating (PAS=1), SOA (Short SOA), or identification accuracy (Incorrectly identified). Two additional tests were contingent on observing evidence for the null hypotheses in the Unconscious and Mostly Unconscious categories – in this case, to test if there was any evidence for disambiguation regardless of the Exposure condition in the U and MU categories, changes in accuracy and meaningfulness would be evaluated Pre to Post-Exposure, on trials pooled across the Catch and Test conditions.

Table 1. Main and null hypotheses, and corresponding analyses fully detailed for Hypothesis 1 contrasting condition U between catch and test trials. The 'a' and 'b' notations refer to the experimental and null hypothesis, respectively. The same approach applied to H2-4 with respect to MU-C. The analyses refer to measures of the two-tones only.

	Main hypothesis	Null hypothesis	Measure	Analysis	Purpose
H1 a/b	The change in **accuracy** following exposure to greyscale images falling in the U category is higher in the test trials than in the catch trials	The change in **accuracy** following exposure to greyscale images falling in the U category is **not** higher in the test trials than in the catch trials	Accuracy	Bayes paired t-test	Test for **objective** disambiguation from exposure to fully unconscious greyscale images
H1 c/d	The change in **meaningfulness ratings** following exposure to greyscale images falling in the U category is higher in the test trials than in the catch trials	The change in **meaningfulness ratings** following exposure to greyscale images falling in the U category is **not** higher in the test trials than in the catch trials	Meaningfulness rating	Bayes paired t-test	Test for **subjective** disambiguation from exposure to fully unconscious greyscale images

2.3 Experiment 2 – Methods

Experiment 2 aimed to replicate Experiment 1, using a 5-alternative-forced-choice task instead of free-naming at the two-tone image identification stage. This method removed all requirements for an experimenter to rate participants' response, which was a potential source of variability. However, correct answers were more likely to be due to chance (chance accuracy of 20%), making the categorisation between the proposed four consciousness categories less straightforward. It was also expected that changes in accuracy from Pre to Post-Exposure would be higher than in Experiment 1, since seeing the correct basic-level category label for a two-tone image, either alone or among distractors, can provide cues for disambiguation (Samaha et al., 2018). Nevertheless, the same would be true following exposure to the catch trials, hence not affecting the validity of the analyses. Unless otherwise mentioned, the same details outlined above applied.

While the two methods of collecting accuracy were distinct, it was expected that the overall pattern of results between Experiments 1 and 2 to be the same – i.e., if a differential effect was observed between test and catch images that were unconscious (Level U) in Experiment 1, the same result was expected in Experiment 2. Having a secondary method of testing accuracy allowed testing the robustness of the effect. Nevertheless, observing an effect in one experiment but not the other would not invalidate the results of either study, but would be interpreted as failure to generalize the findings and would highlight the impact that methodological choices have on studying consciousness.

2.3.1 Participants

A new sample of 60 participants who did not complete Experiment 1 were recruited, following the same protocol in Experiment 1. 12 participants were excluded from all analyses for failing one or more attention checks (5 had accuracies in the visible attention checks under 88%, 7 rated too many visible attention checks with PAS ratings under 3, and 4 rated too many blank attention checks with PAS above 1). The final sample was n = 48 (age range 18-32, mean = 20.8, SD = 3.65, 11 males, 34 females, 3 other identities). All participants had normal or corrected-to-normal vision. No trials were removed because of frame drops.

2.3.2 Materials

To assess participants' identification of the content of both two-tone and greyscale images, a list of 33 correct descriptions of maximum three words (one for each two-tone/greyscale image pair, and 9 attention checks), and a list of 137 plausible but incorrect distractor words (33x4 distractors plus 5 choices for the blank trials), were used. The distractors were matched between themselves and with the correct answers subjectively for visual similarity, and for frequency, as determined by online searches in the British National Corpus Online Services (British National Corpus Consortium, 2007). The list of words was manually collated and validated. To ensure appropriate frequency matching, some descriptors were used as distractors for more than one image, but correct answers were not repeated.

2.3.3 Design

Instead of free-naming, an alternative-forced-choice task appeared, showing the correct answer and 4 distractors. The order of the five options was randomized on each trial.

2.3.3.1 Accuracy

Accuracy was automatically computed from participants' answers, and no answer validation was necessary.

2.4 Pilot experiments

2.4.1 Pilot 1 – Disambiguation effect after conscious images

Pilot 1 was conducted as a positive control, to assess that this paradigm can result in a disambiguation effect when the corresponding greyscale images are presented consciously, for 2 seconds and unmasked. While Chang et. al (2016) showed that greyscale images can lead to disambiguation after a single 16.7ms exposure, it was important to ascertain that any potential failure to replicate the effect is not due to the present stimulus set and paradigm not being able to lead to disambiguation at all. 22 participants were recruited from Cardiff University's participant pool and tested in a group setting. Two participants were excluded due to technical issues, final n = 20 (19 females, mean age 20.17). 24 two-tone/greyscale image pairs were used: half of the two-tones were followed by the corresponding greyscale, and the remaining half by catch greyscale images showing different pictures of semantically

related scenes, i.e., attracting the same name but visually as distant as possible. One image was included to achieve equal two-tone block lengths but was then removed. Participants saw blocks of 8 image pairs, with the Pre- and Post-Exposure stages interleaved (same as in the main experiment, as described in Figure 3 and Figure 4). To reduce the chance identification level and in-lieu of attention checks in the main experiments, in each block were introduced 3 additional greyscale images with no associated two-tones. After each two-tone, participants rated the meaningfulness of each image, and completed a free-naming identification test. Greyscale trials were followed only by the identification test.

For each participant, the change from Pre to Post-Exposure for accuracy and meaningfulness was computed, and Bayesian paired t-tests were used to assess whether the changes were higher in the main trials than catch trials (see Figure 6 for all values). For main trials, there was very strong evidence for an increase from Pre- to Post-Exposure for identification accuracy and meaningfulness ratings. For catch trials, there was strong evidence for an increase in identification accuracy but not in meaningfulness ratings, as expected if the change in accuracy is mainly driven by guessing based on recently viewed pictures rather than a genuine perceptual disambiguation. When comparing the mean changes Pre-Post in the test and catch trials, there was very strong evidence for higher identification accuracy and meaningfulness ratings, comparable across the objective and subjective measures. Altogether, this pattern of results suggests that access to the original template images for 2s provided true perceptual disambiguation.

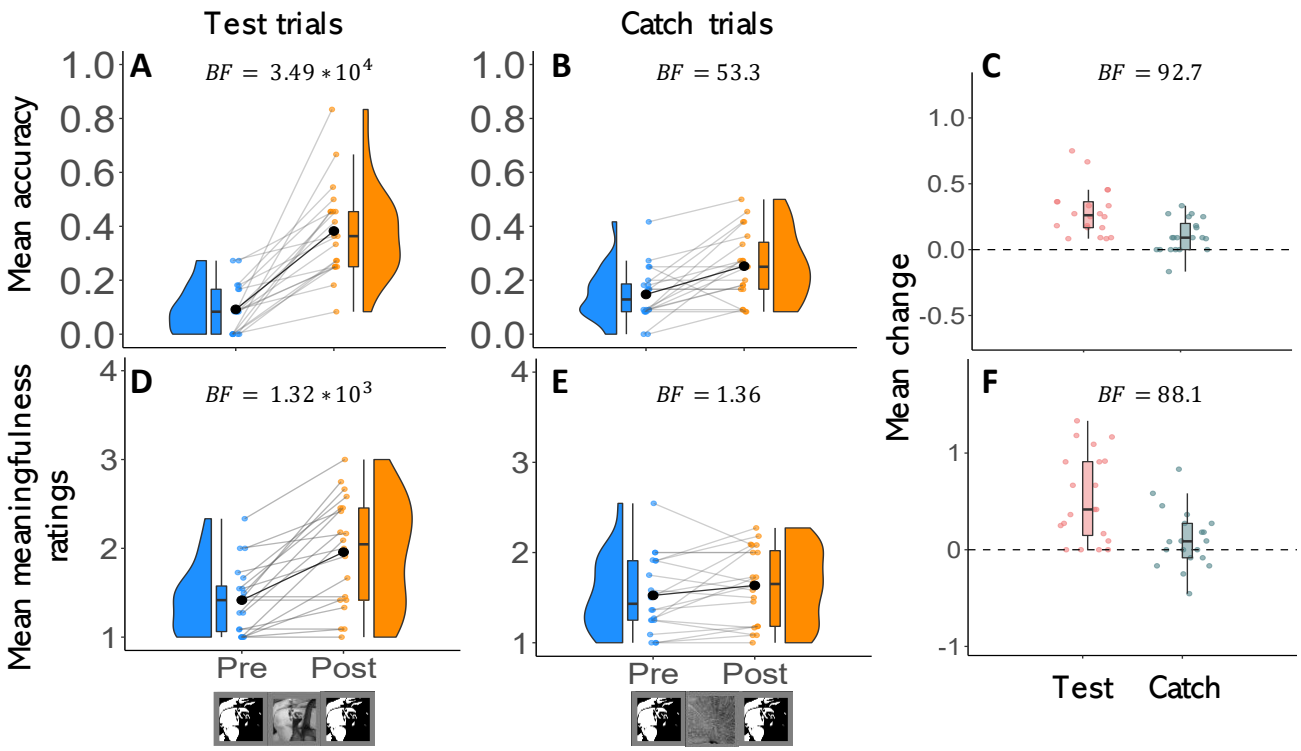


*Figure 6. Distribution of mean accuracy (panels A and B) and meaningfulness ratings (panels D and E) Pre- and Post-Exposure to test and catch trials in Pilot 1. Solid black line represents means of each condition. Panels C and F show the distribution of the changes in the two measures between Pre- and Post- Exposure to test and catch trials. Boxplots show median and IQR, with whiskers representing the minimum, respectively maximum value in the data -/+ 1.5*IQR. Bayes Factor (BF) values are displayed for each comparison, obtained from Bayesian paired t-tests.*

2.4.2 Pilots 2 and 3 – Demonstration of experimental manipulation and trial categorisation

To demonstrate that the planned experimental designs resulted in the expected distribution of datapoints across the four consciousness categories, a set of pilot data was collected for Experiments 1 and 2. The participant recruitment process and experimental design were the same as presented above, apart from catch trials which featured fully unrelated images and 6 catch trials per participant rather than 12. Participants were recruited from both Cardiff University's participant pool (n = 18 for Pilot 2, n = 16 for Pilot 3) and Prolific (n = 14 for Pilot 2, n = 19 for Pilot 3). The same participant exclusion criteria outlined above were followed, leaving n = 24 for Pilot 2 and n = 21 for Pilot 3 after pooling across the two samples, since

they produced similar data. 7 participants were removed for technical errors, 10 for failing at least one attention check, and 5 for failing to report at least a glimpse of an experience (PAS 2 or above) in more than half the trials in the long SOA condition. Trials were divided in the four consciousness categories described above. For both Pilot 2 (Figure 7A) and 3 (Figure 7B), the experimental manipulation resulted in suitable distributions of experiences across the four data bins, allowing testing the hypotheses.

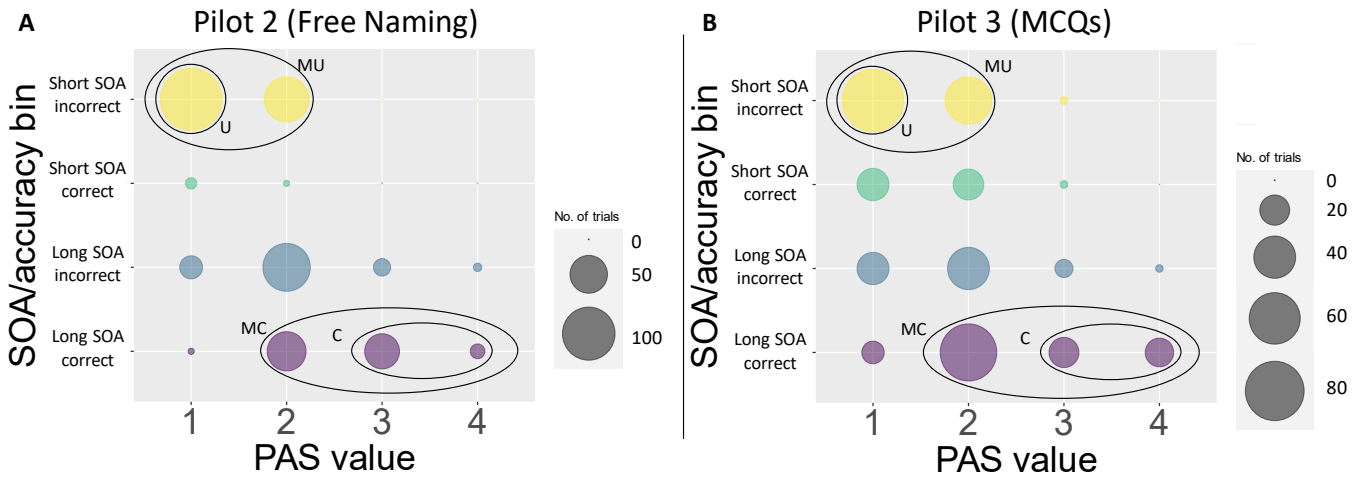


Figure 7. Distribution of trials in each of the four data bins, in A. Pilot 2, and B. Pilot 3. The experimental manipulation resulted in the expected pattern of trials, also showing that participants used the PAS accordingly.

2.5 Experiment 1 results – planned analyses

2.5.1 Combined measurements classification

Table 2 details the results of each Bayesian t-test, in relation to the relevant hypotheses. As shown, there was weak evidence against the hypothesis that accuracy increases more in Test than Catch trials, both under strict (Unconscious category) and less strict (Mostly Unconscious) definitions of 'unconscious'. The same pattern was obtained for meaningfulness ratings, however here the evidence against the alternative hypothesis was strong.

2.5.2 Results-contingent tests for Pre-Post Exposure

Given that there was evidence for the null hypotheses (H1b/H1d/H2b/H2d), the comparisons between Pre- and Post-Exposure for trials pooled across Test and Catch were conducted. For accuracy, there was evidence against an increase from Pre- to Post-Exposure, moderate in the U condition (BF_{null} = 7.79) and weak in the MU condition (BF_{null} = 3.18 ± 0.01%). For meaningfulness, there was weak evidence against an increase in the U condition (BF_{null} = 4.38), but moderate evidence for an increase in the MU condition (BF_{alt} = 5.96). Altogether these results suggest that identification accuracy was not impacted by exposure to the templates/greyscales, irrespective of how strict the criteria for 'unawareness' were. However, counting 'Brief Glimpses' as 'unawareness' resulted in an increase in subjective meaningfulness but not accuracy.

2.5.3 PAS-only classification

Next, Test and Catch trials for two-tones whose corresponding greyscale images were rated as PAS 1 or "No experience" were compared. There was inconclusive evidence for an increase in accuracy, but moderate evidence against an increase in meaningfulness (Table 2).

2.5.4 SOA-only classification

For trials in the Short SOA category, there was weak, respectively strong evidence that accuracy and meaningfulness did not increase in the Test, compared to Catch trials (Table 2).

2.5.5 Accuracy-only classification

Finally, for two-tones whose corresponding greyscale images were incorrectly identified, again there was inconclusive evidence against an increase in accuracy, but strong evidence against an increase in meaningfulness in the Test, compared to Catch trials (Table 2).

*Table 2. Results from the planned analyses in Experiment 1. 'Change' refers to Post minus Pre scores. + and blue text = moderate evidence for the alternative, ++ and blue text = strong evidence for the alternative. * and orange text = moderate evidence for the null, ** and orange text = strong evidence for the null. Text with colour only and no label indicate weak BFs. No label and black text mark inconclusive BFs. All errors were under 0.2%.*

	Hypothesis	Measurement	Mean (SD) change – Test	Mean (SD) change – Catch	Type	BF	Hypothesis supported
Combined measurements classification	H1a/b - U	Accuracy	0.01 (0.216)	-0.019 (0.128)	Null	4.03	Weak – H1b
	H2a/b - MU	Accuracy	0.025 (0.124)	-0.001 (0.127)	Null	2.94	Weak – H2b
	H3a/b – MC	Accuracy	0.172 (0.25)	0.074 (0.185)	Alt	2.05	Inconclusive
	H4a/b – C	Accuracy	0.199 (0.369)	0.064 (0.14)	Alt	1.63	Inconclusive
	H1c/d - U	Meaningfulness	0.007 (0.439)	0.072 (0.465)	Null	14.5**	H1d
	H2c/d - MU	Meaningfulness	0.066 (0.388)	0.117 (0.39)	Null	14.9**	H2d
	H3c/d – MC	Meaningfulness	0.299 (0.595)	0.088 (0.356)	Alt	3.05	Weak – H3c
	H4c/d – C	Meaningfulness	0.409 (0.615)	0.163 (0.382)	Alt	1.29	Inconclusive
PAS-only classification	H5a/b – PAS1	Accuracy	0.029 (0.175)	-0.026 (0.12)	Alt	1.08	Inconclusive
	H5c/d – PAS1	Meaningfulness	0.058 (0.406)	0.049 (0.457)	Null	8.39*	H5d
SOA-only classification	H6a/b – Short	Accuracy	0.026 (0.136)	0.003 (0.116)	Null	3.5	Weak – H6b
	H6c/d – Short	Meaningfulness	0.08 (0.382)	0.121 (0.338)	Null	14.2**	H6d
Accuracy-only classification	H7a/b – Incorrect	Accuracy	0.038 (0.11)	-0.002 (0.121)	Null	1.11	Inconclusive
	H7c/d – Incorrect	Meaningfulness	0.068 (0.304)	0.094 (0.352)	Null	13**	H7d

2.6 Experiment 2 results – planned analyses

2.6.1 Combined measurements classification

Table 3 and Figure 9 detail the results of each Bayesian t-test, in relation to the relevant hypotheses. As shown, for the key tests (U and MU), there was moderate to strong evidence against the hypothesis that accuracy increases beyond catch trials, neither under the strictest definition of 'unconscious' (Unconscious category) nor under a less strict definition in the Mostly Unconscious category.

2.6.2 Results-contingent tests for Pre-Post Exposure

Given that there was evidence for the null hypotheses (H1b, H1d, H2b, H2d), the comparison between Pre and Post for trials pooled across Test and Catch was conducted. For accuracy, there was weak evidence for an increase in both the Unconscious condition (BF_{alt} = 5.08) and the Mostly Unconscious condition (BF_{alt} = 3.63). For meaningfulness ratings, there was strong to extremely strong evidence for an increase in both the Unconscious condition (BF_{alt} = 44.4) and the Mostly Unconscious condition (BF_{alt} = 810).

2.6.3 PAS-only classification

Next, Test and Catch trials for two-tones whose corresponding greyscale images were rated as PAS 1 or "No experience" were compared. There was strong evidence that accuracy and meaningfulness did not increase in the Test, compared to Catch trials (Table 3).

2.6.4 SOA-only classification

For trials in the Short SOA category, there was moderate, respectively strong evidence that accuracy and meaningfulness did not increase in the Test, compared to Catch trials (Table 3).

2.6.5 Accuracy-only classification

Finally, for two-tones whose corresponding greyscale images were incorrectly identified, there was strong evidence that accuracy and meaningfulness did not increase in the Test, compared to Catch trials (Table 3).

Table 3. Results from the planned analyses in Experiment 2. The same notation convention applies as in Table 2. All errors were under 0.2%.

	Hypothesis	Measurement	Mean (SD) change – Test	Mean (SD) change – Catch	Type	BF	Hypothesis supported
Combined measurements classification	H1a/b - U	Accuracy	0.1 (0.343)	0.107 (0.44)	Null	8.17*	H1b
	H2a/b - MU	Accuracy	0.06 (0.249)	0.093 (0.322)	Null	13**	H2b
	H3a/b – MC	Accuracy	0.113 (0.457)	-0.023 (0.309)	Alt	1.26	Inconclusive
	H4a/b – C	Accuracy	0.046 (0.554)	0.062 (0.46)	Null	8.24*	H4b
	H1c/d - U	Meaningfulness	0.226 (0.475)	0.501 (0.828)	Null	22.6**	H1d
	H2c/d - MU	Meaningfulness	0.174 (0.342)	0.223 (0.447)	Null	13.9**	H2d
	H3c/d – MC	Meaningfulness	0.407 (0.576)	0.06 (0.48)	Alt	21.4++	H3c
	H4c/d – C	Meaningfulness	0.583 (0.881)	0.044 (0.57)	Alt	4.68	Weak – H4c
PAS-only classification	H5a/b – PAS1	Accuracy	0.077 (0.298)	0.1 (0.373)	Null	10.4**	H5b
	H5c/d – PAS1	Meaningfulness	0.176 (0.438)	0.329 (0.806)	Null	18.07**	H5d
SOA-only classification	H6a/b – Short	Accuracy	0.63 (0.199)	0.047 (0.258)	Null	6.54*	H6b
	H6c/d – Short	Meaningfulness	0.149 (0.275)	0.158 (0.301)	Null	10.14**	H6d
Accuracy-only classification	H7a/b – Incorrect	Accuracy	0.025 (0.215)	0.063 (0.247)	Null	15.2**	H7b
	H7c/d – Incorrect	Meaningfulness	0.159 (0.26)	0.209 (0.39)	Null	15.1**	H7d

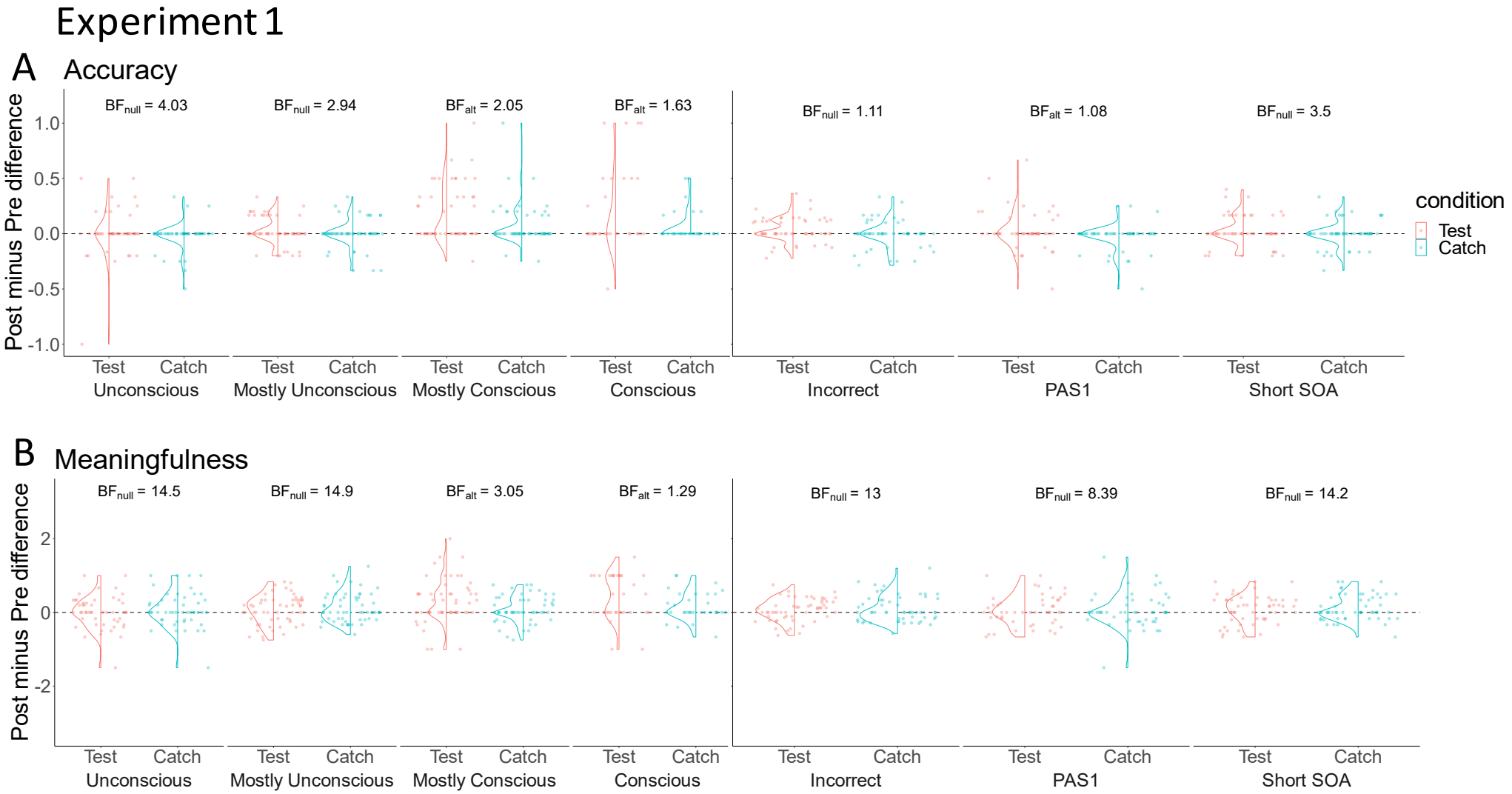


Figure 8. Experiment 1 results, for all classifications, for accuracy (Panel A) and meaningfulness ratings (Panel B). Each dot is a participant's change in means between Pre- and Post-Exposure (Post minus Pre). The dashed line marks 0 (no change). The subscript next to each BF marks whether the evidence favoured the alternative hypothesis ("alt", Test higher than Catch) or the null ("null", Test not higher than Catch).

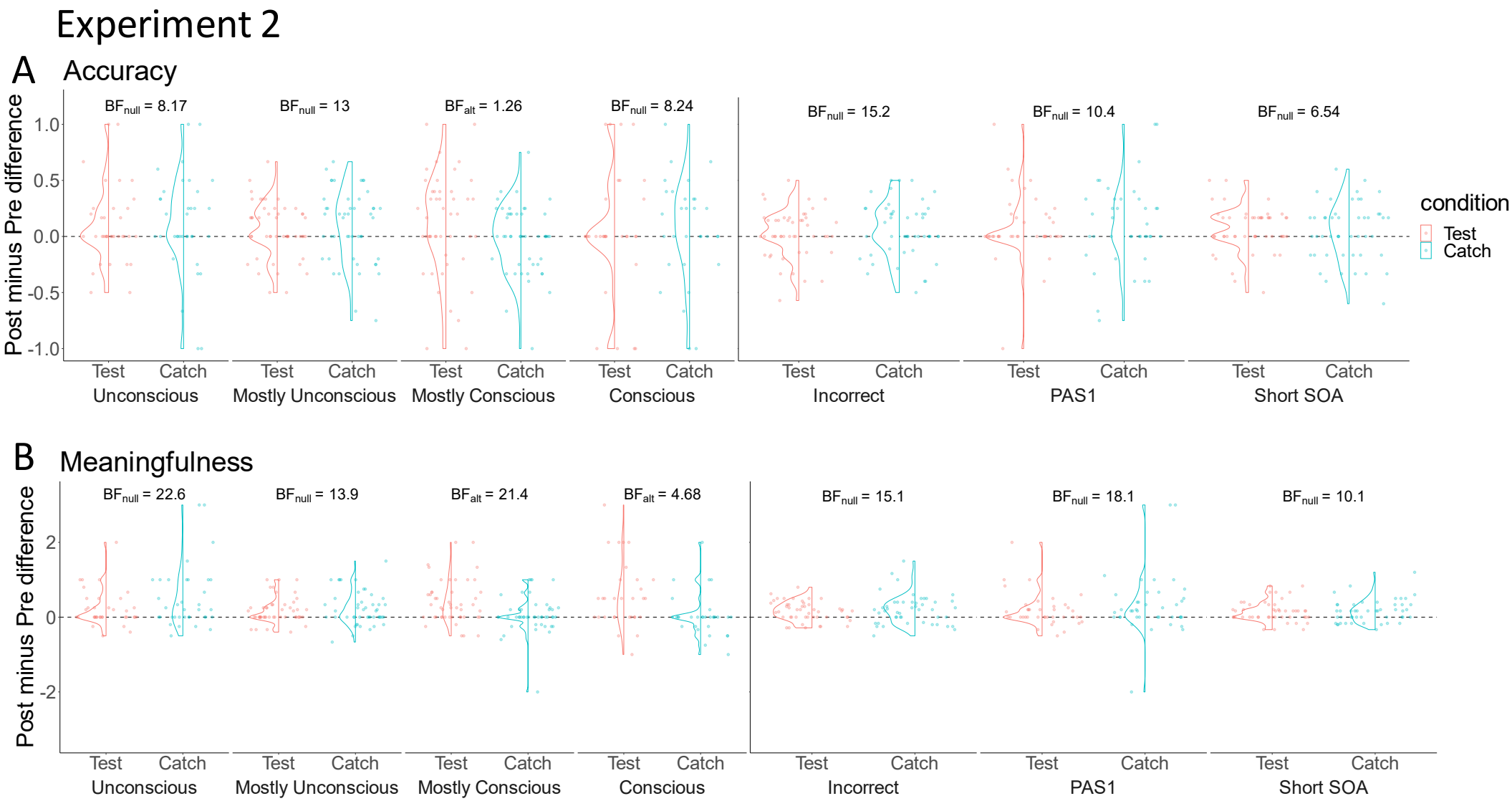


Figure 9. Experiment 2 results, for all classifications, for accuracy (Panel A) and meaningfulness ratings (Panel B). The same details apply as in Figure 8.

2.7 Exploratory analyses

2.7.1 Condition-dependent Pre-Post Exposure changes

As explained above, a related question is whether there was evidence of disambiguation (Pre-Post) at all in the two unconscious conditions (U and MU). This question was addressed in 2.5.2 and 2.6.2, with trials from Test and Catch trials pooled together. However, the weak evidence against a difference between the Test and Catch trials in Experiment 1 (2.5.1, Table 2) might be seen as not fully justifying pooling them together. Therefore, the Pre-Post comparisons in 2.5.2 were repeated without pooling, i.e., separately for the Test and Catch trials. These analyses were not needed for Experiment 2, since there the evidence against a difference in the Unconscious/Mostly Unconscious conditions was stronger. The same parameters applied as in previous tests ('wide' r-scale, one-tailed). These analyses map onto Figure 8, comparing each distribution to 0 (dashed line). Results from all comparisons are included in Table 4. The evidence favours the null hypothesis that in neither condition the increase in accuracy was different from 0, under the strictest definition of 'unawareness' (Unconscious condition), however the evidence was less conclusive for the Mostly Unconscious Test condition. For Meaningfulness ratings, the evidence was again less conclusive, but it overall favoured the null, except for the Catch condition.

Table 4. Results from exploratory analyses on changes between Pre- and Post-Exposure in Experiments 1 and 2. Only the MC and C conditions are included in Experiment 2, because the strong evidence for the null in both measures did not justify condition-dependent analyses. The same notation convention as before applies.

		Test		Catch	
		Accuracy	Meaningfulness	Accuracy	Meaningfulness
Experiment 1	Unconscious	BF_{null} = 6.77 *	BF_{null} = 8.21 ±0.12% *	BF_{null} = 17.4 **	BF_{null} = 2.92 ±0.03%
	Mostly Unconscious	BF_{null} = 1.66	BF_{null} = 2.45	BF_{null} = 10.2 ±0.05% **	BF_{alt} = 2.25
	Mostly Conscious	BF_{alt} = 6.03 × 10^3 ++	BF_{alt} = 83.7 ++	BF_{alt} = 10.9 ++	BF_{null} = 1.01
	Conscious	BF_{alt} = 16.4++	BF_{alt} = 95 ++	BF_{alt} = 5.51	BF_{alt} = 3.87
Experiment 2	Mostly Conscious	BF_{null} = 1.15	BF_{alt} = 3.07 × 10^3 ++	BF_{null} = 12.6**	BF_{null} = 3.85
	Conscious	BF_{null} = 4.82	BF_{alt} = 65.1 ++	BF_{null} = 3.58	BF_{null} = 4.96

The same comparisons were repeated for the Mostly Conscious and Conscious conditions in Experiments 1 and 2, to assess if there is evidence that conscious exposure leads to disambiguation. In 2.5.1 and 2.6.1, the evidence weakly or inconsistently supported a difference between Test and Catch, thus again not justifying pooling. As shown in Table 4, the evidence supports the hypothesis of an increase from Pre to Post Exposure, in both markers of disambiguation, when participants reported awareness of the images during Exposure, but only in Experiment 1.

2.7.2 Bayes factors for the effect from Chang et al. (2016)

As explained before (2.2.6.2), Chang and colleague's (2016) key result comes from the significant difference in comparing disambiguation in Test and Catch trials. For identification accuracy, the reported test result was t = 2.076, p = 0.0492, n= 24 (page 7), however for subjective recognition, the test statistic was not reported. To assess what the Bayes Factor for the identification accuracy comparison would have been, the t value and n were entered into the *ttest.tstat* function in the BayesFactor package (Morey et al., 2022), which allows estimating the BF from the test statistic of paired t-tests. Because it was not specified by Chang and colleagues if this test was one or two-tailed, and to assess the robustness of the conclusion under different scales of the prior, 4 tests were computed, none of which indicated any robust evidence for the alternative hypothesis: one-tailed "medium" (BF_{alt} = 2.55), one-tailed "wide" (BF_{alt} = 2.07), two-tailed "medium" (BF_{alt} = 1.32), two-tailed "wide" (BF_{alt} = 1.06). Therefore, these analyses indicate that at best, the evidence for this effect is inconclusive; in other words, the evidence from Chang and colleague (2016) does not support the conclusion that two-tone identification accuracy was different following exposure to unconscious templates than following exposure to a blank screen.

2.7.3 Condition-independent Pre-Post Exposure changes in single measurement classifications

In some (Experiment 1, 2.5.4) or all (Experiment 2, 2.6.3-2.6.5) the single measurement classification analyses, there was evidence against higher increases in the Test condition compared to Catch. While the evidence was weak or inconclusive in accuracy comparisons in Experiment 1, altogether these results suggest that regardless of how 'unconscious' is

defined, there seems to be no added benefit of exposure to the corresponding template image. However, it could still be the case that there was disambiguation from Pre to Post-Exposure. To test for this possibility, mirroring the planned results-contingent analyses (2.5.2 and 2.6.2), Test and Catch trials were pooled together and assessed for an increase from Pre to Post-Exposure, separately for accuracy and meaningfulness, for all three single measurement classifications. Results are included in Table 5.

Altogether, these results suggest that using a single measurement for assessing unawareness in this design can lead to substantially inconsistent findings about whether there is an increase in disambiguation (i.e., Pre to Post-Exposure). Moreover, how disambiguation is defined (at the subjective or objective level) and how objective accuracy is tested (MCQs or Free Naming) can further lead to discrepancies in conclusions.

Table 5. Pre-Post Exposure comparisons, pooled across Test and Catch trials, for the single-index classifications of unawareness. The same notation convention as in Table 2 applies.

	Experiment 1		Experiment 2	
	Accuracy	Meaningfulness	Accuracy	Meaningfulness
Incorrect only	$BF_{alt} = 1.61$	$BF_{alt} = 3.3$	$BF_{null} = 1.43$	$BF_{alt} = 3.87 \times 10^{3\,++}$
PAS1 only	$BF_{null} = 7.51^{*}$	$BF_{null} = 3.4$	$BF_{alt} = 3.38$	$BF_{alt} = 14.2^{++}$
Short SOA only	$BF_{null} = 2.94$	$BF_{alt} = 15.6^{++}$	$BF_{alt} = 2.85$	$BF_{alt} = 1.28 \times 10^{3\,++}$

2.7.4 Enhanced exclusions criteria

One exclusion criterion the preregistered protocol did not anticipate is how to handle means in each data bin based on only one or few images. Such values can be seen as unreliable and could add noise to the distributions. To explore whether the key results in 2.5.1 and 2.6.1 (U and MU conditions) would be robust when these participants were removed, the main Test-Catch comparisons in both Experiments 1 and 2 were recomputed, excluding from each comparison participants based on whether their means in either the Test or the Catch trials were based on fewer than 2, 3 or 4 trials. The BFs for each comparison can be found in Table 6. As observed, in Experiment 1 there is no evidence for a higher increase in Test compared

to Catch, regardless of how the trials are parsed. Moreover, for the key analyses that initially yielded inconclusive BFs (Experiment 1 Accuracy, U and MU conditions, Figure 8), increasing the number of trials per mean generally strengthened the evidence for the null hypothesis. This analysis lends further credibility to the explanation that altogether in this dataset the evidence favours the null hypothesis, although the low number of images introduces variability in the strength of the evidence. Figure 10 shows the same distributions as Experiment 1 and 2 U and MU conditions, with symbols indicating in which of the analyses below participants were included.

Table 6. Results from Test vs Catch comparisons for U and MU conditions, analogue to Tables 2 and 3, manipulating the reliability of each mean per participant. Experiment 2 MU 2+ trials BFs are the same as in Table 3, because all means were based on at least 2 trials. The same notation convention as in Table 2 applies. All errors were under 0.2%.

Experiment	Measurement	Condition	2+ trials	3+ trials	4+ trials
Experiment 1	Accuracy	Unconscious	BF_{null} = 1.22	BF_{null} = 2.61	BF_{null} = 9.72 *
		Mostly Unconscious	BF_{null} = 3.89	BF_{null} = 3.85	BF_{null} = 7.94 *
	Meaningfulness	Unconscious	BF_{null} = 10.7 **	BF_{null} = 13.1 **	BF_{null} = 8.72 *
		Mostly Unconscious	BF_{null} = 14.8 **	BF_{null} = 14.6 **	BF_{null} = 17.4 **
Experiment 2	Accuracy	Unconscious	BF_{null} = 7.44 *	BF_{null} = 5.12	BF_{null} = 7.22 *
		Mostly Unconscious	BF_{null} = 13 **	BF_{null} = 7.77 *	BF_{null} = 10.3 **
	Meaningfulness	Unconscious	BF_{null} = 15.7 **	BF_{null} = 11.3 **	BF_{null} = 8.05 *
		Mostly Unconscious	BF_{null} = 13.9 **	BF_{null} = 11 **	BF_{null} = 18.5 **

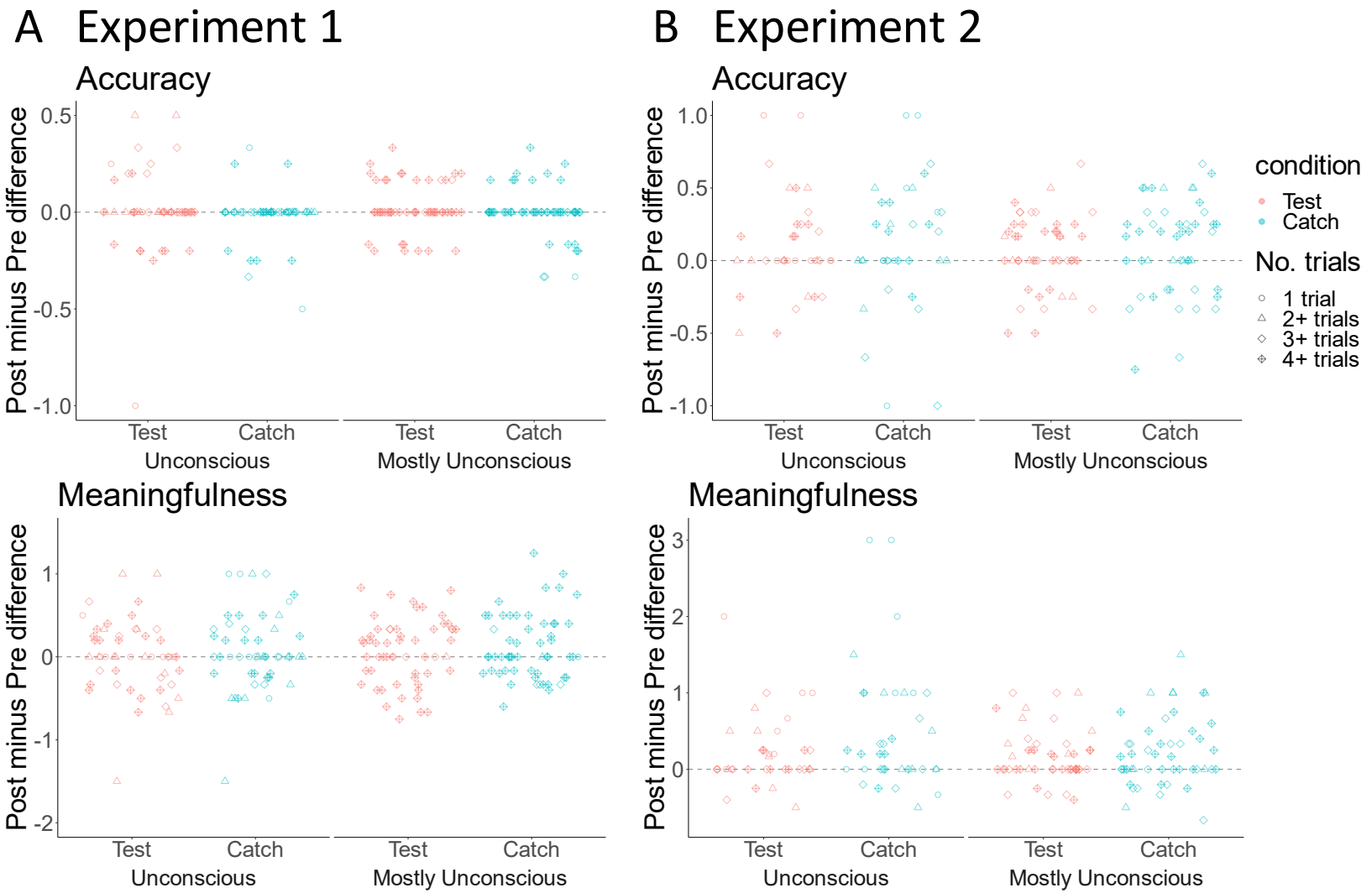


Figure 10. Re-plotting of datapoints from the key conditions in Figures 8 and 9, indicating the number of trials each mean was based on. The scales were adjusted between figures to maximize visibility.

2.8 Discussion

In this set of experiments, participants were exposed to complex natural images (templates/greyscales) with different levels of visibility. Before and after exposure to these images, participants were exposed to two-tone stimuli that were either directly derived from the images (Test trials), or that only depicted the same content as the images (Catch trials) – usually, two-tone stimuli can only be perceived as meaningful (i.e., disambiguated) after exposure to the template natural images. The main goal of the experiments was to reassess the finding that disambiguation of two-tone stimuli can follow exposure to template images that participants had no awareness of (Chang et al., 2016). A secondary goal was to assess the extent to which different criteria for post-hoc selecting 'unconscious' trials impact conclusions about the presence of an effect.

The study did not replicate the finding from Chang et al. (2016) that identification accuracy increased beyond the Catch condition, following exposure to 'unconscious' template/greyscale images. In both Experiments 1 (2.5.1) and 2 (2.6.1), using two definitions of 'unawareness' (a strict Unconscious condition, and a more liberal Mostly Unconscious condition), there was no evidence for a higher increase in the Test, compared to Catch group. Instead, the evidence favoured the null hypothesis, albeit weakly in Experiment 1. These findings are complemented by the exploratory analysis (2.7.2), which found that the BF for the key identification comparison in Chang and colleagues' study also does not show robust evidence for a difference, despite a p-value below the cutoff point of 0.05. For the subjective marker of disambiguation, the findings from meaningfulness ratings are consistent with the pattern reported by Chang and colleagues: strong Bayesian evidence against the hypothesis that increases in the Test condition were higher than in Catch, again in both Experiments 1 and 2 and across two definitions of 'unawareness'. Altogether, these results converge towards the conclusion that there is no added benefit of exposure to the corresponding templates when there is no awareness of the images during the Exposure stage (under a classification based on multiple indices).

Regarding the secondary goal of comparing awareness labelling criteria, there was again no evidence for a higher increase in Test compared to Catch following exposure to unconscious grey-scale images, in either Experiments, measurements, or any of the three classifications used to defined (based on Exposure stage SOA only, PAS only, or Accuracy only; sections

2.5.3-2.5.5 and 2.6.3-2.6.5). For most tests (all comparisons in Experiment 2 and Meaningfulness tests in Experiment 1), the evidence moderately or strongly favoured the null hypothesis of no higher increase. For Experiment 1 accuracy, there was only weak evidence for the null in the SOA-only classification, the other comparisons remaining inconclusive. It could be argued that the reason why the PAS-only and Accuracy-only conditions produced inconclusive results for Accuracy in Experiment 1 was because of noise in the data. The SOA-based classification was the only condition that did not rely on trial selection based on participants' responses, so each mean per participant was based on the highest number of trials possible, thus reducing the noise around the mean. This explanation does not apply for meaningfulness ratings, where there was strong evidence for the null in all three classifications despite being based on the same trials as the accuracy tests; however, it could be that one measure is generally noisier than the other. Altogether, these findings suggest that, while overall there is no criterion of 'unawareness' under which there is evidence of an advantage of Test compared to Catch, experimental design choices (measurement, criterion) can still impact the strength of the evidence.

Interestingly, there was also no consistent evidence for an advantage of Test images compared to Catch in the conscious conditions. In Experiment 1, the evidence only weakly favoured an advantage for both measures (Figure 8). In Experiment 2, the evidence was mixed, with both support for the null (Conscious condition, accuracy) and for an increase (Mostly Conscious condition, meaningfulness, Figure 9). Follow-up analyses (2.7.1) found a disambiguation effect (Pre- vs Post-Exposure) in the conscious conditions in Experiment 1 (all but one comparison), but only in Meaningfulness Test trials in Experiment 2. Moreover, while the pattern of findings seems consistent within studies, they substantially differ between studies - suggesting that the disambiguation effect might not be robust, and highlighting that different experimental design choices could lead to different conclusions.

As for whether there was Pre-Post Exposure disambiguation at all when pooling across Test and Catch (2.5.2 and 2.6.2), the evidence is more mixed. In line with the expected differences (2.3), the method for probing identification accuracy impacted the results: when there was no exposure to the labels in Experiment 1 (2.5.2), the evidence favoured the null hypothesis of no increase in accuracy above 0, in either unawareness conditions, however in Experiment 2 there was evidence for an increase above 0. For meaningfulness, there was

extremely strong evidence for an increase in Experiment 2 in both unawareness conditions, and moderate evidence for an increase only in the Mostly Unconscious condition in Experiment 1. Follow-up exploratory analyses for Experiment 1 (2.7.1) showed that even when testing Pre-Post disambiguation separately for the Test and Catch conditions, there was evidence that no disambiguation occurred in either measure when using the strictest definition of 'unawareness'. Moreover, the only evidence for an increase was weak, only in the Catch condition, under a less strict definition of 'unawareness', and only in the subjective marker of disambiguation. Thus, this not consistent with Chang and colleague's findings, and suggests counterintuitively that exposure to a catch grayscale led to better subjective disambiguation than the relevant template. In any case, a few things could explain the increases from Pre to Post-Exposure in the absence of a difference between conditions. A first theoretically possible explanation is spontaneous disambiguation, i.e., simply from viewing the two-tone image a second time. Secondly, it could be that unconscious Test and Catch images both evoked the correct two-tone category, and that this semantic priming contributed to a genuine perceptual disambiguation of the Post-Exposure two-tone images. It is also inevitable that two different images of the same category share some visual features (e.g., stripes for a zebra), that may be sufficient to induce a perceptual disambiguation even when unconscious. In the Chang et al.'s Catch condition, these two image-based factors were absent and could therefore not contribute to genuine disambiguation, however in the current study's design they could. Therefore, the current design of these experiments does not allow disentangling between possible explanations for the increases between Pre- and Post-Exposure, in the U and MU categories in Experiment 2. Another possible explanation, relevant to Experiment 2 only, is that participants had access to additional 'hints' about the image content through the labels, which could have led to disambiguation (consistent with Samaha et al., 2018). Indeed, this would explain why the Pre-Post increases were observed primarily in Experiment 2 compared to Experiment 1.

While it is not fully warranted therefore to conclude that information from unconscious template/greyscale images played no role in disambiguation, it is unlikely that this increase was due to any unconscious image contribution. First, a Pre-Post increase was observed primarily in Experiment 2, despite the image exposure being the same as in Experiment 1. As mentioned above, it is thus plausible that the increase was due to exposure to the labels in

the MCQ options, which was the only difference in design with Experiment 1. Secondly, it would be difficult to interpret Pre-Post increases in Meaningfulness alone (Experiment 1 MU condition, section 2.5.2) as reflecting real disambiguation in the absence of an increase in identification accuracy, since they could be due to perceived demand bias, with participants recognizing they had seen the two-tone before and rating it as more meaningful on the second presentation. Moreover, observing it only in the MU condition (PAS at least 'brief glimpse', Short SOA, incorrect identification of the template content) and not the U condition (PAS only 'no experience', same other criteria as MU) might also be because participants having at least a brief glimpse of the template content might have been enough to lead to a subjective feeling of meaningfulness. This explanation would be aligned with previous evidence - albeit from objective tasks – that performance was higher when participants reported 'brief glimpses' compared to 'no experience' (e.g., Overgaard et al., 2004; Ramsøy & Overgaard, 2004; Sandberg et al., 2010).

The study does have limitations. One design limitation, mentioned above and explored in additional analyses (2.7.4), is the low number of images in the experimental set – consequently the low number of trials that each mean per participant is based on. This possible issue is more pronounced in the combined measurements than in the single measurement classifications, since each trial had to pass multiple exclusion criteria. The exploratory analyses lend some support for this interpretation - progressively strengthening the exclusion criteria (by only including means based on a certain number of trials) also strengthens the evidence for the null in the critical U and MU conditions. However, this limitation does not affect the validity of the analyses, because the low number of images was compensated for with a higher number of participants. Nevertheless, future work following these findings could benefit from having a higher number of two-tone and template image pairs, that had undergone similarly extensive piloting and refinement as the current stimulus set (Teufel et al., 2015). Finally, both methods of judging two-tone recognition have limitations: the subjectivity of rating Free Naming answers in Experiment 1, and introducing exposure to the correct labels in Experiment 2, that can contribute to disambiguation. While both aspects affected in a similar way Test and Catch conditions in each respective experiment, these considerations do complicate the interpretation of the Pre-Post changes. A different approach that future research could use is a grid-test, where

participants are asked to select from a list of pre-defined locations on the image where different elements might be (e.g., the eyes in a face image, Ludmer et al., 2011). This technique would fully bypass the issue of subjectivity in Experiment 1, and introduce only minimal and non-descriptive 'hints' about the content. However, it would not be suitable for assessing identification of masked, briefly-presented templates, thus introducing a substantial task difference between the Exposure stage and the rest of Pre-Post stages.

Another future direction, to disentangle between spontaneous and category exposure-related disambiguation, would be to conduct a follow-up experiment, in which half of two-tone images would be associated with the Catch greyscales (the current Catch condition in both experiments), and the other half would not be associated with any images during the Exposure stage (i.e., no images would be presented). In this experiment, firstly, it would be expected to find the same pattern of Pre-Post changes in the Catch trials Experiment 2, and the catch trials in the follow-up experiment, across all categories, since the experimental conditions would be virtually identical. Furthermore, observing a difference between conditions in the follow-up experiment would be evidence that semantic information, exposure to the labels, or the category-specific low-level visual information contributed to the Pre-Post disambiguation observed in Experiment 2. Alternatively, observing evidence against a difference between conditions would be evidence that the Pre-Post effect in Experiment 2 was fully driven by spontaneous disambiguation, while conscious exposure to the labels, unconscious semantic information and category-specific visual information played no role.

Nevertheless, the present experiments highlight a few conclusions. Regarding unconscious one-shot learning, when controlling for key confounds and making more conservative the criteria for 'unawareness', the evidence does not support previously published conclusions. Regarding methods of studying unconscious effects, these results suggest that the chosen objective measure can impact conclusions. The choice of criteria for 'unawareness' can also, in some circumstances, lead to different conclusions (e.g., that the results are inconclusive, or that they support the null) – although here they were mostly robust. To allow better comparison across findings, future work would benefit from systematically comparing conclusions about specific research questions from a wider variety of experimental approaches.

Chapter 3

3.1 Introduction

Successful adjustment to environmental conditions is a cornerstone of survival. To achieve this, the brain allows information that it previously encountered to shape the processing of new inputs. As explained in Chapter 1, this susceptibility to the influence of past inputs has been thoroughly observed in the sensory domain, with one example being perceptual learning, or the long-term changes in subjective experience and performance that occur adaptively based on demands from the environment (Gibson, 1969; Sagi, 2011). Such effects have been reliably induced in the laboratory through extensive practice on a task, demonstrating the existence of complex neural plasticity in the adult brain. In the visual domain, visual perceptual learning (VPL) has been observed in a variety of tasks, from detection and discrimination of high-level stimuli such as faces from noise, to more low-level visual features such as orientation and luminance contrast (for reviews, see Fine & Jacobs, 2002; Watanabe & Sasaki, 2015).

Nevertheless, the neural mechanisms and the necessary conditions of learning are yet to be fully mapped. A series of studies (Schwiedrzik et al., 2009, 2011) proposed that VPL can develop from stimuli that are initially below the threshold of objective discrimination. In their experiments, Schwiedrzik and colleagues first exposed participants to multiple trials in which a simple shape (square or diamond) flashed on the screen for 10ms, between (Schwiedrzik et al., 2009) or only followed (Schwiedrzik et al., 2011) by a mask. The interval between the stimulus onset and the subsequent mask onset (stimulus onset asynchrony, or SOA) was systematically manipulated, to find the SOA which yielded a sensitivity (i.e., d-prime) not significantly different from 0 in discriminating between shapes. Subjective awareness, as measured by the Perceptual Awareness Scale (PAS; Ramsøy & Overgaard, 2004) was also collected. Then, participants completed 5 training sessions (3000 trials in total) only on the chosen SOA, during which the same two measures (Schwiedrzik et al., 2011) or the discrimination measure only (Schwiedrzik et al., 2009) were collected. Both papers reported that d-prime sensitivity, as well as mean PAS ratings at the chosen SOA increased from before to after training. Moreover, this transition to d-primes significantly

above 0 was reported to occur in the first 100-200 trials of the first training session (Schwiedrzik et al., 2009). This finding led the authors to conclude that the initially indiscriminable stimuli broke into awareness with practice, and that this effect occurred very early in the training.

These claims are worth exploring further. Since publication, these papers (Schwiedrzik et al., 2009, 2011) cumulated over 130 citations (57, respectively 74, Google Scholar, 4/12/2023), with relevance for different fields in consciousness research. Specifically, the findings have been discussed in the light of multiple theories of consciousness (ToCs), due to their implications for what the role(s) or function(s) of awareness might be. Schwiedrzik and colleagues' findings are frequently cited as consistent with theories that posit that consciousness is something that we learn to do, by the proponents of these theories (the 'radical plasticity thesis' as in Timmermans et al., (2012); the self-organizing metarepresentational account – SOMA – as in Cleeremans et al., 2020), as well as others (e.g., Siedlecka et al. 2020). For instance, Timmermans et al. (2012) suggest that *"metacognition […] is an active, trained construction process"*, which they link with Schwiedrzik et al. (2009)'s findings, taken to *"support the idea that one can train people to gain conscious access to their own representations"* (p. 1418). In the same vein, Cleeremans et al., (2020) recently proposed that *"consciousness should be viewed as a process that results from continuously operating unconscious learning and plasticity mechanisms"* (p. 112), and Schwiedrzik et al. (2009) findings are cited in support to the central claim that *"perception is continuously shaped by learned priors"* (p. 115). Although Schwiedrzik and colleagues' findings have been used to illustrate the general idea underlying these theories (i.e., that training can result in gaining awareness of visual stimuli that were previously not consciously experienced), it remains unclear whether these theories necessarily and specifically predict these findings, and what the implications would be for these theories should they not be observed.

On the other hand, other ToCs predict that learning can only occur if the stimuli to learn from are consciously experienced (Baars & Franklin, 2007; Kugele & Franklin, 2021; Lamme, 2006, 2010, 2014; Meuwese et al., 2013). Under the recurrent processing theory (RPT), Lamme and colleagues argue that *"[t]he function of conscious vision may be to add a final layer to our interpretation of the world, to solve relatively "new" visual problems, and to*

enable visual learning" (Lamme, 2014, p. 1), implying that unconscious visual stimuli cannot enable visual learning. Similarly, the global workspace theory (GWT), suggests that only once information is consciously experienced in the GW can it be further distributed to brain areas for complex cognitive processing, including memory, inner speech, and *'almost all kinds of learning'* (Baars, 2005, p. 47). Baars, Franklin, and colleagues further proposed the Learning Intelligent Distribution Agent (LIDA, Baars & Franklin, 2007; Kugele & Franklin, 2021; Ramamurthy et al., 2006) as a model of GWT, and the Conscious Learning Hypothesis that *"all significant learning is evoked by conscious contents, but the learning process itself and its outcomes may be unconscious"* (Baars & Franklin, 2007, p. 957)*,* with PL being one of the four types of learning under this hypothesis. Albeit disagreeing on the underlying brain mechanisms, both lines of reasoning would therefore predict no behavioural effects of learning from unconscious stimuli – prediction challenged by Schwiedrzik and colleagues' findings, as indeed acknowledged but not addressed by Lamme (2014). Altogether, while Schwiedrzik and colleagues' findings cannot ultimately help discern between ToCs, they have ramifications for the study of consciousness and the current complex theoretical landscape.

The present study aimed to expand upon Schwiedrzik and colleagues' findings, and retest the claim that objective and subjective sensitivity to visual information change due to prolonged practice. Instead of comparing individual SOA levels, this paradigm models, under metacontrast masking, each participant's contrast discrimination and detection psychometric functions (PFs). PAS ratings were also collected throughout all sessions. The reasons to manipulate contrast instead of SOA are two-fold: first, to circumvent the issue of individual differences in the link between SOA and performance in metacontrast masking (i.e., type A showing a linear function of accuracy increasing with SOAs, or type B showing a U-shaped function, Albrecht et al., 2010), which could introduce difficulties in analysing and interpreting the results. Secondly, there is converging evidence from human and animal studies that training-induced improvements occur in both contrast detection (Bao et al., 2010; Furmanski et al., 2004; Sowden et al., 2002; Yu et al., 2016) and discrimination (Hua et al., 2010; Scholes et al., 2021), albeit from paradigms with different stimuli. For example, Sowden and colleagues (2002) conducted a 2 interval-forced-choice task (2IFC), in which a grating was presented in only one of two consecutive time intervals, over 10000 trials. The gratings had fixed orientation, were not masked, and the contrast level was fixed at the

value that generated 70.7% accuracy in a pre-training session for each participant (hence above values that would be considered reliably unconscious). The ability to correctly detect in which interval the gratings appeared improved over training, to an average of ~79% accuracy.

In the current study, a contrast level yielding under 60% discrimination accuracy (Day 1) was chosen for each participant. Participants in the Learning group then trained for one session of 1000 trials at the chosen contrast (Day 2). A separate Control group did not complete this training. In Day 3, all participants completed the PF measurements again. If extended practice on a stimulus unlikely to be consciously experienced leads to VPL, as suggested by Schwiedrzik and colleagues, then shifts in the PFs towards lower contrast levels would be expected - meaning that the same contrast level would generate higher performance in Day 3 compared to Day 1. Increases in mean PAS would also be expected, with both effects being larger for the Learning group than the Control group.

As mentioned above, the current paradigm also includes a detection measure, besides discrimination. While it is not uncommon to use discrimination either as index for learning (e.g., Furmanski et al., 2004; Hua et al., 2010; Scholes et al., 2021) or awareness (e.g., Nishina et al., 2007; Schlaghecken et al., 2008), detection has also been used for both purposes (learning index e.g., Bao et al., 2010; Furmanski et al., 2004; Sowden et al., 2002; Yu et al., 2016, awareness index: e.g., Balsdon & Azzopardi, 2015; Heeks & Azzopardi, 2015). The relationship between detection and discrimination in masking paradigms is not well-explored, although many consciousness researchers reported in a survey to believe that discrimination thresholds are lower than detection thresholds while also believing that this pattern has not be convincingly demonstrated in the literature (Peters & Lau, 2015). However, Schwiedrzik and colleagues (2009; 2011) used the same measure, discrimination d-prime, both to assess unawareness and to measure learning – besides PAS as solely an index of learning. This approach limits the generalizability of the effect to paradigms that use detection as an index of awareness. It also limits the conclusions that can be drawn about a learning effect to discrimination only. Therefore, despite not factoring detection into the selection of the stimulus parameters for training, it is still beneficial to collect it, for two reasons. First, it allows assessing if there are any differences pre-training in discrimination and detection (i.e., if conclusions drawn about initial ‘unconsciousness’ of the stimuli based

on discrimination would align with those based on detection). Secondly, it allows testing whether training on a discrimination task with a stimulus below the discrimination threshold would produce comparable learning outcomes in a detection measure, or whether the conclusions about learning are indeed limited to discrimination only.

Four planned analyses were therefore conducted, answering each of the following questions:

- Q1P. Does discrimination sensitivity increase more in the Learning group compared to the Control group?
- Q2P. Does detection sensitivity increase more in the Learning group compared to the Control group?
- Q3P. Does subjective visibility improve more in the Learning group compared to the Control group?
- Q4P. Is there a difference between discrimination and detection sensitivity, before and after training?

The differences in experimental designs between the present study and Schwiedrzik and colleagues' might limit the comparability of the findings. To minimise these differences, exploratory analyses aligned to the one they performed were conducted, on d-prime values (Q1E-Q2E). Given that d-prime are standardized measures and hence task-independent, the changes in d-prime in the present study were compared to the changes in d-prime over a comparable number of trials from Schwiedrzik and colleagues' studies (Q3E, 3.4.4).

3.2 Methods

3.2.1 Participants

51 participants, naïve to the purpose of the experiment, were recruited from Cardiff University School of Psychology in exchange of course credit or payment. All participants had normal or corrected-to-normal vision. Due to the fast-paced presentation of the stimuli in the experiment which could appear flicker-like for some individuals, participants who had a history with photosensitivity and/or epilepsy were not eligible. Participants were informed that participation in Days 2 (if in the Learning group) and 3 was contingent on their accurate

performance on Day 1, and that they could also earn a bonus of £2/session for good performance. In practice, all participants that completed the experiment received the bonuses. 2 participants were excluded for non-completion. 9 participants completed only Day 1 because at least one of their bootstrapping tests (n = 1000, parametric if goodness-of-fit p-value was over 0.05, non-parametric if below or equal 0.05, Kingdom & Prins, 2016) resulted in a failed fit with a pattern best explained by a constant function (i.e., inflection point tending towards infinite), meaning that no inflection point confidence interval could be computed. Data from 1 participant was excluded after completion because the bootstrapping failed during preprocessing, suggesting that the initial CI estimate was not reliable.

3.2.2 Materials

The experiment was run in a laboratory setting, and was displayed on a Dell P2213 monitor with a diagonal of 22 inch, a resolution of 1680x1050, and a refresh rate of 60Hz. Brightness and contrast were set to 75. The computer had an integrated graphics card (Intel HD Graphics 4600), 8-bit depth and standard dynamic range. Participants were seated at 44cm away from the screen, and head movements were restricted by a chin and forehead rest. The experiment was custom written and run in MATLAB 2019b, using the Psychophysics Toolbox extensions, version 3.0.18 (Brainard, 1997; Pelli, 1997). Stimuli images were created using Microsoft PowerPoint and consisted of pairs of arrows pointing to the left ("<<"), subtending 3.5cm by 2cm, or 4.55 x 2.60 degrees of visual angle (DVA). The right-pointing equivalent and the mask (Figure 11) were obtained through custom MATLAB (*MATLAB*, 2021) code from the left-arrow image. The metacontrast mask consisted of a frame of 16 overlapped pairs of arrows at maximum contrast (Figure 11), subtending 10.4cm x 5.9cm, or 13.5 x 7.67 DVA. The PAS (Ramsøy & Overgaard, 2004) was used, with four steps: (1) No experience, (2) A brief glimpse, (3) An almost clear experience, and (4) A clear experience (Sandberg et al., 2010). The description of the scale was modified from its original form, to remove mentions of confidence and focus strictly on the clarity of the experience (see Chapter 2 for a similar approach). The full description of the scale as well as the quiz items used in training can be found in Appendix 4. The contrast levels were defined in relation to the range of grey shades possible to display on the monitor, such that 256 = white, 128

background grey, and 0 = black. Here, the contrast levels are reported as indices, with 0 meaning no contrast (blank). The monitor was not gamma corrected. A lookup table for the associated cd/m^2 values is available in Appendix 5.

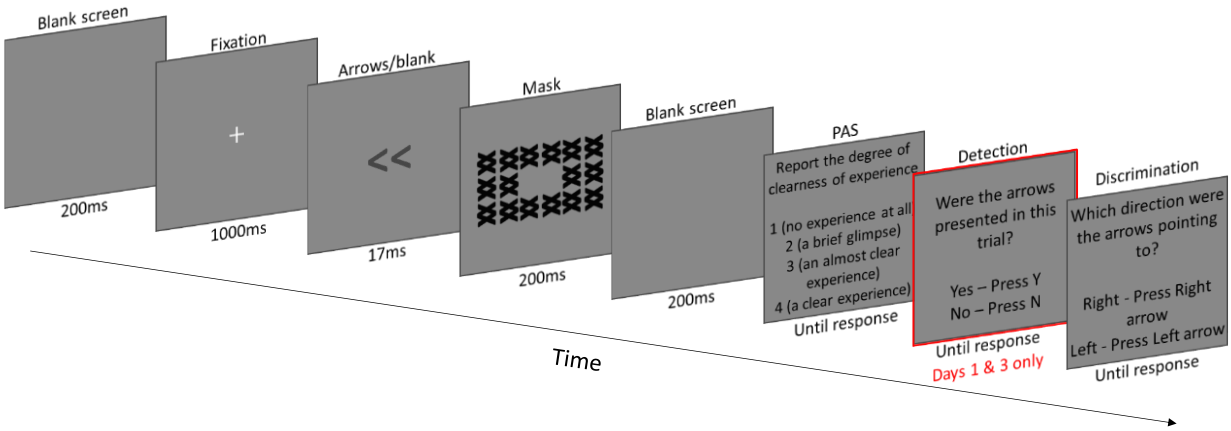


Figure 11. Time sequence of a trial, common across all sessions. The ratio of stimuli to screen size is larger than in the experiment for illustration purposes. The arrows contrast varied across trials and participants. The Detection question was only present in Days 1 and 3.

3.2.3 Design and procedure

The experiment took place over three days. Day 1 began with a presentation of the stimuli and task, a self-paced study period of the PAS, and then a short quiz verifying participants' understanding of the different scale levels. The quiz repeated until they reached 7/8 (87%) correct. The quiz was followed by a practice run consisting of 14 trials (4 blank, 10 fully visible arrows), where participants had trial by trial feedback for both detection and discrimination. Then, participants completed 14 blocks of the method of limits (MoL), to quickly determine their optimal range of contrasts (13 levels) spanning from chance to ceiling (more details in Appendix 5). Due to an error, 1 participant completed only 6 MoL trial blocks (3 ascending, 3 descending). The number of MoL blocks was chosen so as to be sufficiently high to give a good estimate of the range while not substantially lengthening the experiment, with the number of blocks being comparable to other studies (e.g., 12 in Hock & Schöner, 2010). The range of contrasts included in the MoC task had maximum granularity, meaning that intermediary levels were not possible – which was important to maximize the chances of having data at the training contrast, and hence allow comparisons outside of the

PFs. Such comparisons allow drawing parallels with other studies that do not use PFs (e.g., Schwiedrzik et al., 2009). The size of the range was chosen keeping in mind that non-experienced observers might have lower consistency in their answers; hence, using a more restricted range would increase the risk of not capturing the critical information for deriving the PFs. Then, they completed a method of constant stimuli (MoC) task, where 36 trials at each of the 13 levels were shuffled and presented among 60 blank trials (total of 528 trials/participant). The relatively high number of trials per contrast level was again chosen so that there was sufficient data for subsequent analyses on individual contrast levels (performance as well as PAS answers). The MoC task was split across 6 blocks separated by self-paced breaks up to 1 minute. No feedback was given. The same trial specifications as in Figure 11 applied. Reaction times were not logged. A buffer of 200ms was introduced between the answer on each task and the onset of the next screen. At the end of Day 1, PFs were plotted on the detection and discrimination data. Based on the discrimination PF only, the standard error of the inflection point was calculated, using bootstrapping (see details in 3.2.1 and Kingdom & Prins, 2016). Then, another PF was fitted, with the same parameters but a modified inflection point (initial inflection point contrast minus SE), from which was selected the highest contrast value that yielded under 60% accuracy. The rationale was to maximize the intensity of the stimulus that could be presented without reliable discrimination performance (for a similar labelling of under 60% performance as 'failed perception', see Karni & Sagi, 1993). Discrimination was chosen as an index of awareness for consistency with the previous studies (Schwiedrzik et al., 2009; 2011). However, for one participant, the detection PF was used, because while the discrimination and detection PFs were near-identical, a CI could only be computed for the detection PF. They were subsequently removed from PF-based analyses (Q1P, Q2P, Q4P – see section 3.3.1 for details). For participants in the Learning group, a value predicting 95% accuracy was also chosen, to be used in the visible attention checks in Day 2.

On Day 2, participants in the Learning group completed 1180 trials (1000 experimental, 40 visible attention checks, 40 blank attention checks, 100 blank), distributed over 10 blocks separated by self-paced breaks up to 1 minute. The trial structure was identical to that on Day 1, except that the detection question was removed (Figure 11). The attention checks also included a 500ms ISI between the arrows and the mask. At the end of each block

participants were told the percent discrimination accuracy in the experimental trials of the blocks they have just completed. Incorrect answers to visible attention checks were followed by a short reminder to pay more attention; no similar feedback was given for PAS answers. Participants in the Control group did not complete any session on Day 2. In Day 3, all participants completed the MoC task, with the same parameters as in Day 1. The sessions in Days 1 and 2 lasted around 75 minutes, while Day 3 lasted around 40 minutes.

3.2.4 Data cleaning and analysis

All preprocessing and statistical analyses were conducted in RStudio v2021.9.1.372 (RStudio Team, 2021) using R v4.2.1 (R Core Team, 2021) and the following packages: BayesFactor v0.9.12-4.4 (Morey et al., 2022), tidyverse v2.0.0 (Wickham et al., 2019; Wickham & RStudio, 2023), data.table v1.14.2 (Dowle et al., 2021), gghalves v0.1.4 (Tiedemann, 2022). Additionally, the 'dprime' function in the package psycho v0.6.1 (Makowski, 2018) was used. Trials with frame gains (n = 14) were removed from all analyses. Fitting of the PFs was conducted in MATLAB v2021a (*MATLAB*, 2021) using the Palamedes toolbox v1.10.9 (Prins & Kingdom, 2018). Given that the x-axis was linear and 0 meant absence of stimulus, Weibull functions were fitted, with a fixed guess rate (0.5) and lapse rate (0.01), and free parameters for inflection point and slope, for each participant and Days (1 and 3).

To account for possible biased answers in the detection question, the raw detection scores at each contrast level were adjusted, using the following formula:

Equation 1.

$$\text{prop}_{\text{HitsUnbiased}} = \frac{(num_{Hits}/(num_{ContrastTrials})+(1-(num_{FalseAlarms}/(num_{Blank}))}{2}$$

to obtain an unbiased proportion of Hits at each contrast level. This proportion was then multiplied by the total number of trials at each contrast level. All detection PFs were computed on these estimates.

The analyses consisted of Bayesian equivalents of within and between-subjects t-tests (Q1P-Q4P) and correlation (Q4P). Details about the BayesFactor package are included in Chapter 2 (2.2.6). All tests used a prior with the scale of 1 ('wide' label for t-tests, 'ultrawide' label for ANOVA), meaning that for t-tests the expected differences between compared groups (i.e., effect size) were up to 1. A directional prior was also included, to specifically test for whether performance was better in the Learning compared to Control, and Day 3 compared to Day 1. This directional prior could be negative in the case of inflection points (i.e., decreases from Pre-Post meaning improved performance), or positive in the case of mean PAS. The directionality of each analysis is specified in each section.

Analyses are split into quality checks, planned analyses (question number followed by a P) which are the most relevant for this experiment's design and which target the key questions in the introduction, and exploratory analyses (question number followed by an E), focusing on links with previous literature or additional questions worth exploring that were not planned for. For Q1P and Q2P, improvements are defined as reductions in the PF inflection points, and conversely increases in accuracy. For Q3P, improvement is defined as an increase in mean PAS.

3.3 Results – planned analyses

All BF errors (i.e., error of the underlying BF estimation, indicating that the real value is between BF ± error, Doorn et al., 2019) were under 0.01%, unless otherwise specified. Table 7 contains mean and SDs for each variable, group, and session. Values are rounded to three significant digits.

3.3.1 Quality checks and exclusions

The chosen exclusion criteria (EC) were:

1. 90%+ of visible attention checks (Day 2) discriminated correctly;
2. 90%+ of visible attention checks (Day 2) answered with PAS2 or above;
3. 70% of blank attention checks (Day 2) answered with PAS1;
4. 70% of blank normal trials (all days) answered with PAS1;
5. No bootstrapping tests finding scenarios best explained by a constant function, for any of the four PFs (discrimination/detection, Day 1/ Day 3).

EC1 is justified because the contrast was calibrated from the Day 1 PF to yield accuracy close to ceiling. EC2 was on par with EC1 following the assumption that awareness influences both accuracy and PAS ratings, so it is expected that high awareness as indicated by accuracy should also entail at least PAS 2 ("Brief Glimpse") responses, if participants used the PAS as intended. The cutoff of 70% of blank trials answered with PAS 1 in ECs 3 and 4 is justified by the reliable findings that the proportion of stimulus-absent trials answered with PAS 1 is much lower than ceiling but at least 70% (Chapter 4 Q3, 4.3.3). While it is unclear why such ratings occur, setting the criterion on par with the visible checks at 90%+ would therefore be potentially unachievable. No participants failed EC1. Two participants failed EC2 (answered only 87.5% visible checks with PAS2+), and two other failed EC3, so all four were excluded from all analyses. Nine participants in total failed EC4 (9 in Days 1 and 3, 4 in Day 2), so they were removed from PAS analyses only (Q3P). 11 participants failed EC5 (7 in the Learning group, 4 in the Control group). They were removed only from the analyses involving PF metrics (Q1P, Q2P, Q4P, Q4E).

The discrimination and detection PFs were also compared between the Learning and Control groups at Day 1, using Bayesian independent t-tests with a two-sided prior. There was moderate evidence against a difference in inflection points (discrimination BF_{null} = 3.55 ± 0.01%, detection BF_{null} = 3.69 ± 0.01%) and slope (discrimination BF_{null} = 3.7 ± 0.01%, detection BF_{null} = 1.71 ± 0.01%), therefore suggesting that the groups did not differ prior to the experimental manipulation.

Table 7. Descriptive statistics for the planned analyses Q1P-Q3P, namely inflection points and accuracy at the chosen contrast from the discrimination and detection PFs, and mean PAS across all trials, in Day 1 and Day 3, for both groups.

		Inflection point				Accuracy from PF				PAS rating (all trials)	
		Discrimination		Detection		Discrimination		Detection			
		Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Learning	Day 1	14.9	5.63	16.3	5.22	0.554	0.020	0.553	0.028	1.57	0.27
	Day 3	12.5	4.66	14.1	5.17	0.588	0.074	0.567	0.060	1.70	0.33
Control	Day 1	15.1	3.52	16.3	3.28	0.560	0.010	0.559	0.036	1.50	0.16
	Day 3	12.8	2.78	14.2	2.47	0.609	0.071	0.594	0.065	1.73	0.30

Table 8. Descriptive statistics for the exploratory analyses Q1E-Q2E, namely discrimination and detection sensitivity (d-prime) and mean PAS at the chosen contrast.

		d-prime				PAS ratings chosen contrast	
		Discrimination		Detection			
		Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Learning	Day 1	0.270	0.383	0.155	0.337	1.07	0.078
	Day 3	0.646	0.653	0.550	0.578	1.13	0.126
Control	Day 1	0.153	0.508	0.297	0.468	1.125	0.086
	Day 3	0.375	0.943	0.484	0.595	1.19	0.189

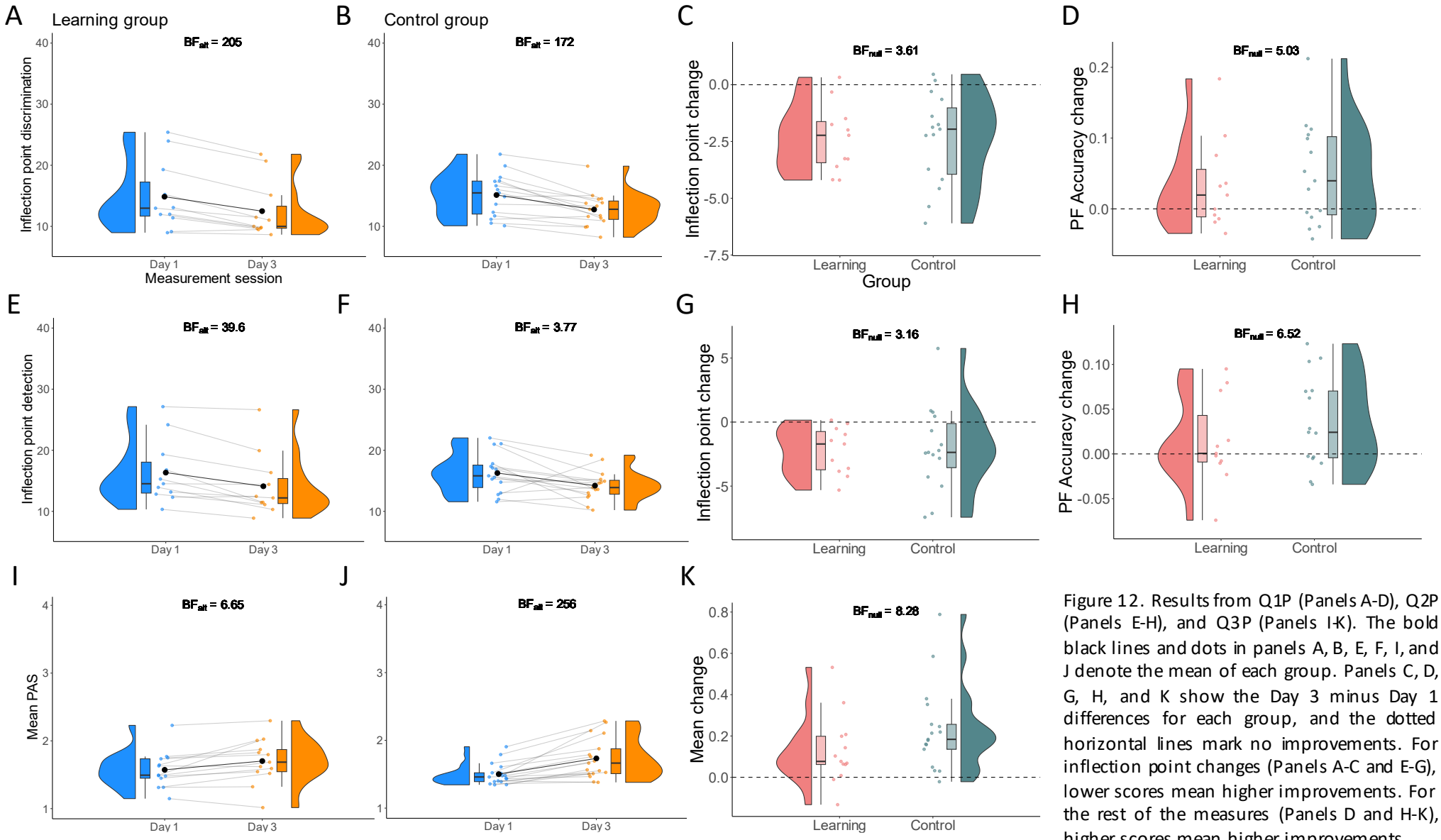


Figure 12. Results from Q1P (Panels A-D), Q2P (Panels E-H), and Q3P (Panels I-K). The bold black lines and dots in panels A, B, E, F, I, and J denote the mean of each group. Panels C, D, G, H, and K show the Day 3 minus Day 1 differences for each group, and the dotted horizontal lines mark no improvements. For inflection point changes (Panels A-C and E-G), lower scores mean higher improvements. For the rest of the measures (Panels D and H-K), higher scores mean higher improvements.

Figure 12.

3.3.2 Q1P. Does discrimination PF shift more in the Learning group compared to the Control group?

To answer this question, discrimination inflection points in Days 1 and 3 were first compared, for the Learning and Control groups separately (Figure 12A & Figure 12B). Means and SDs can be found in Table 7. There was extremely strong evidence for a reduction in both the Learning and Control groups, showing that improvement occurred. However, when comparing the reduction across groups (Day 3 minus Day 1, Figure 12C), there was evidence against these reductions being larger in the Learning compared to the Control groups. A similar pattern was observed for PF accuracy at the chosen contrast (i.e., accuracy on the PF at the chosen contrast), albeit with stronger evidence in the Catch condition (BF_{alt} = 5.74), compared to Test (BF_{alt} = 1.48), and moderate evidence against a higher increase in the Learning group compared to Control (Figure 12D).

3.3.3 Q2P. Does detection PF shift more in the Learning group compared to the Control group?

The same analyses as for Q1P were repeated for detection. There was evidence for a reduction in both the Learning group (Figure 12E), and the Control group (Figure 12F), but evidence against these reductions being larger in the Learning compared to the Control groups (error ± 0.01%, Figure 12G). For PF accuracy at the chosen contrast, again evidence favoured more an increase in the Control group (BF_{alt} = 5.79) than the Learning group (BF_{null} = 1.88), with moderate evidence against a higher increase in the Learning group compared to Control (Figure 12H).

3.3.4 Q3P. Does subjective visibility improve more in the Learning group compared to the Control group?

There was at least moderate evidence for an increase in mean PAS (mean across all contrasts) in Days 1 and 3 in both the Learning (Figure 12I), and the Control group (Figure 12J). As in Q1P and Q2P, there was moderate evidence against the hypothesis that mean PAS increased more in the Learning group compared to Control (Figure 12K).

3.3.5 Q4P. Is there a difference between discrimination and detection sensitivity, before and after training?

One important question is whether discrimination and detection PFs differed before training. Since the training contrast was chosen based on discrimination only, a difference between measures would suggest that should detection have been chosen as an index of awareness, conclusions about learning might have been different. Because no differences were expected nor found (3.3.1) before training between the two groups, data from the Learning and Control groups was collated. A Bayesian paired, two-sided t-test comparing Discrimination and Detection inflection points in Day 1 found evidence for a difference (BF_{alt} = 3.79), with the Detection inflection points higher than Discrimination (Figure 13). However, comparing the accuracies at the trained contrast from PFs in Day 1 showed moderate evidence for the null (BF_{null} = 6.55 ±0.07%). A two-way ANOVA with Measure and Day as predictors on inflection point values and participant as random effects found evidence for both main effects (Day: $BF_{alt} = 2.92 \times 10^5 \pm 0.22\%$, Measure: $BF_{alt} = 17.3 \pm 0.36\%$), but no interaction ($BF_{null} = 6.4 \pm 0.51\%$), consistent with the conclusion that the training did not impact the two measures differently.

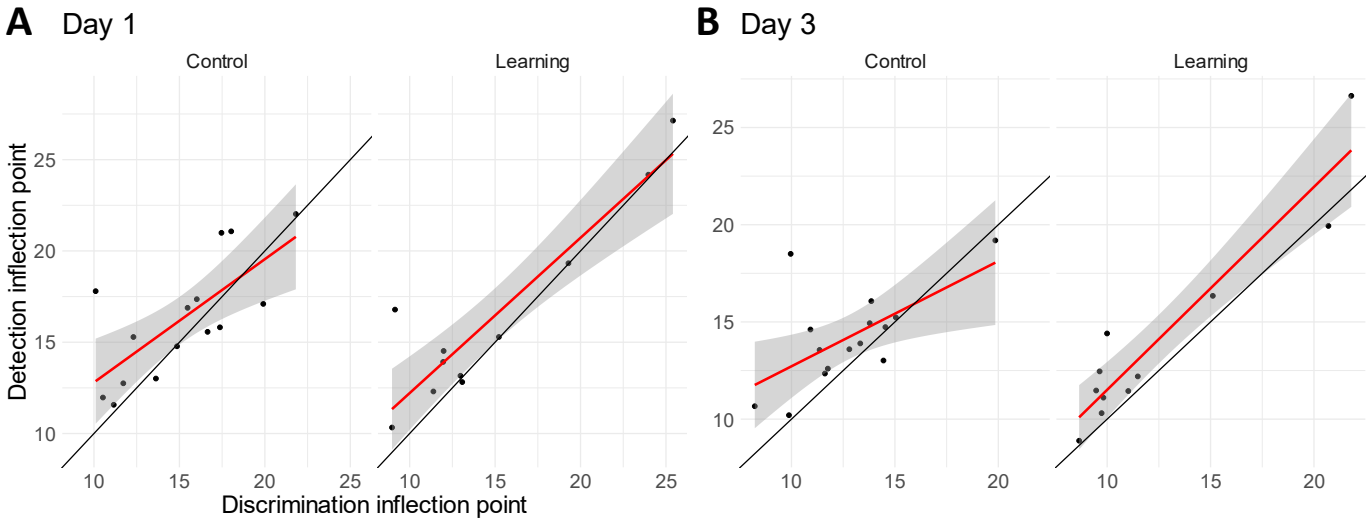


Figure 13. Scatter plot showing the relationship between discrimination and detection inflection points for each group, in Day 1 (Panel A) and Day 3 (Panel B). The black line marks the unity line.

3.3.6 Interim conclusions

To summarize the findings from the planned analyses, there was evidence of no higher increases in the Learning group compared to Control, in none of the three measures tested (discrimination, detection, and subjective experience as indexed by mean PAS), neither for inflection points nor accuracy at the chosen contrast from the PFs. At the chosen contrast, the evidence for performance increases was present in the Control group but inconclusive in the Learning group. This pattern therefore suggests that while learning occurred between Days 1 and 3, it was not driven by the training in Day 2. Furthermore, while there were differences between detection and discrimination PFs before training, they were not present at the lower end of the functions, thus highlighting that using detection instead of discrimination would have led to similar conclusions about awareness.

3.4 Results – exploratory analyses

The exploratory analyses focused on changes in d-prime discrimination and detection sensitivity and mean PAS at the chosen contrast (Q1E), whether there is an increase in discrimination accuracy and mean PAS during the training in Day 2 for the Learning group (Q2E), and how the changes in d-prime in the current experiment compare with changes reported in previous studies (Q3E).

3.4.1 Quality checks and exclusions

No PF-based exclusions were applied for any of the tests here (except in Q4E), because only the raw data is analysed rather than model parameters, however the attention-check based exclusions were applied. For 12 participants, the chosen contrast was lower than the tested range, and therefore they were not included in analyses of trials at the chosen contrast (Q1E, most analyses in Q2E). The same participants removed from Q3P for failing EC4 were removed from PAS analyses only in Q1E and Q2E. For consistency, the same 'ultrawide' prior scale was used in the ANOVA (Q2E, Q4E).

3.4.2 Q1E. Are there any changes in d-prime and mean PAS at the chosen contrast?

Sensitivity (d-prime) and mean PAS only in the trials at the chosen contrast (Table 8) were considered, mirroring the post-hoc tests in Schwiedrzik et al., (2009). For the Learning group, there was at least weak evidence for an increase from Day 1 to Day 3 in all three measures:

discrimination d-prime (BF_{alt} = 2.56), detection d-prime (BF_{alt} = 3.17), and mean PAS (BF_{alt} = 2.72). For the Control group, the evidence was inconclusive: discrimination (BF_{null} = 1.63), detection (BF_{null} = 1.34), and mean PAS (BF_{null} = 1.06). Testing for higher increases in the Learning group than Control also yielded inconclusive support for the null for discrimination (Figure 14A) and detection (Figure 14B), but moderate evidence against a difference in mean PAS (Figure 14C). Altogether, these mixed results highlight qualitative differences in patterns of findings between PF-based analyses and d-prime analyses.

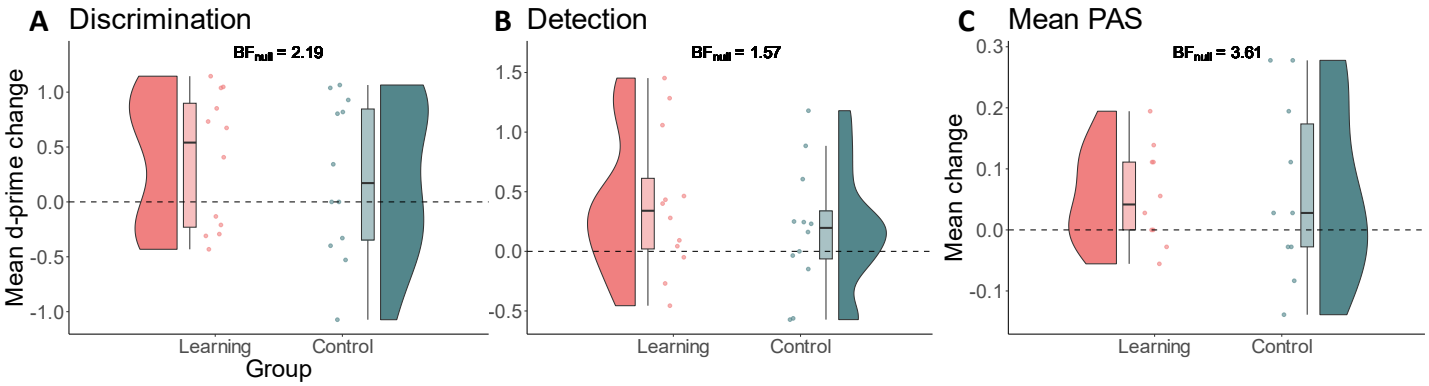


Figure 14. Changes (Day 3 minus Day 1) in mean d-prime (panels A and B) and mean PAS (panel C) for each group, and each measurement.

3.4.3 Q2E. Is there an increase in discrimination accuracy and mean PAS throughout the training session in the Learning group?

This question was deemed as exploratory because it does not directly relate to the comparisons between groups necessary to test the main hypotheses. However, given the overall, albeit weak, support for increases in d-prime between Days 1 and 3 in the Learning group (Q1E), one question to explore further is when during training this change might have occurred. Discrimination d-prime (Figure 15A) and mean PAS (Figure 15C) were computed for each of the 10 blocks (100 trials each), followed by Bayesian equivalents of analyses-of-variance (ANOVA), with block as predictor. There was extremely strong evidence against an effect of Block on d-prime, $BF_{null} = 2.21 \times 10^3$ or mean PAS, $BF_{null} = 1.30 \times 10^3$. Including participants as random effects weakened the evidence for the null but did not change the pattern ($BF_{null} > 100$ for both analyses).

The next tests assessed whether there was an increase in performance from Block 1 to Block 10, and found evidence for the null (i.e., no increase) in both measures (Table 9). For better visualization of the d-prime and PAS progression throughout the experiment, means from Days 1-3 were computed for the subsample of the Learning group where the trained contrast was included in the measured range (d-prime, Figure 15B; mean PAS, Figure 15D) from Days 1-3. For this subgroup, as shown in Table 9, there was evidence for an increase from Day 1 to Block 1 in Day 2 but evidence against an increase between Block 10 and Day 3. These findings thus strongly suggest that the training did not impact either subjective or objective awareness of the arrows, and any increases were due to practice in Day 1.

*Table 9. Results for Q3E. + and blue text = moderate evidence for the alternative, ++ and blue text = strong evidence for the alternative. * and orange text = moderate evidence for the null, ** and orange text = strong evidence for the null. Text with colour only and no label indicate weak BFs. No label and black text mark inconclusive BFs.*

	Day 1 vs Day 2 Block 1 (subgroup)	Day 2 Block 1 vs Block 10	Day 2 Block 10 vs Day 3 (subgroup)
d-prime	$BF_{alt} = 15.7$ ++	$BF_{null} = 4.28 \pm 0.07\%$ *	$BF_{null} = 7.41$ *
Mean PAS	$BF_{alt} = 4.06$ +	$BF_{null} = 4.95 \pm 0.01\%$ *	$BF_{null} = 7.49$ *

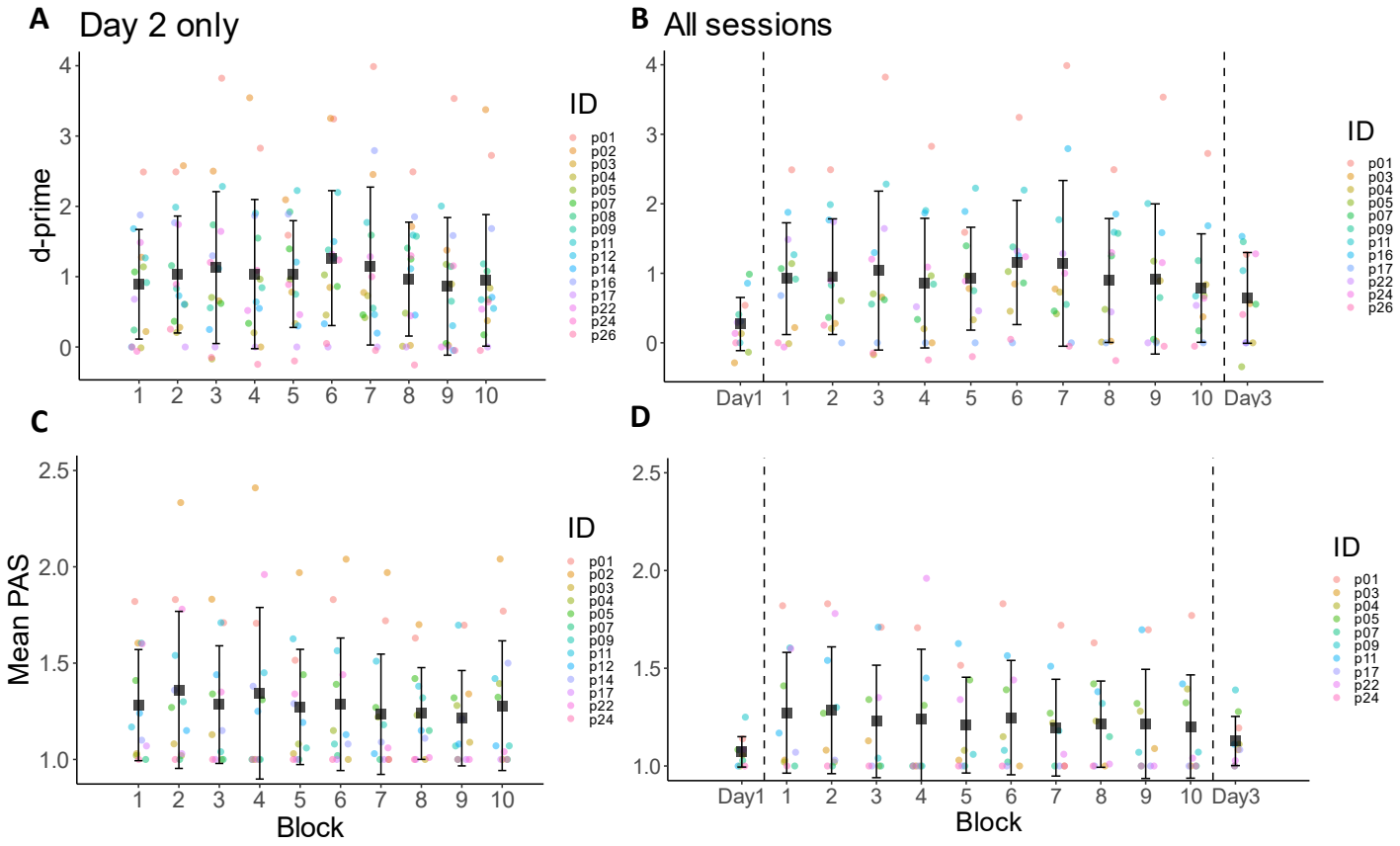


Figure 15. Discrimination sensitivity d-prime (Panels A and B) and mean PAS (Panels C and D) for each block and participant in the training session (Learning group only). Black central rectangles and error bars show the mean of d-prime/means ± 1SD in each block. Panels B and D show values for the subsample of participants from the Learning group with trials at the trained contrast in Days 1 and 3.

3.4.4 Q3E. Does the change in discrimination d-prime at the trained contrast differ from the change in d-prime in Schwiedrzik and colleagues' experiments, for a comparable number of trials?

One possible explanation for the improvements in Q1P and Q3P, indirectly supported by findings from Q2E, is that they were driven by learning during the measurement session in Day 1, which included both sub-threshold and above-threshold contrasts. If this explanation would also apply for Schwiedrzik and colleagues' studies, then their result should show a similar increase in d-prime over a comparable number of trials (up to 1000) as from Day 1 to Day 3 in the current study. Since Schwiedrzik et al. (2011) do not report SD/SEM values for

the mean d-prime at the trained SOA before training or after each session, nor block-by-block descriptive statistics, only the findings from Schwiedrzik et al. (2009) were considered.

Each training session in Schwiedrzik et al. (2009) – equivalent to Day 2 – had 6 blocks of 100 trials, however block-by-block d-prime values were only reported for the first training session. Mean d-prime at the trained SOA, as averaged over all trials in the first threshold measurement session, was reported as 0.067 (SD = 0.335, extracted from SEM in Figure 3A from Schwiedrzik and colleagues, 2009, using WebPlotDigitizer from Rohatgi, 2015), and not significantly different from chance. Mean d-prime in Block 6 (i.e., final block of the first session) was 0.59 (SD = 0.38) for the full sample.

50000 simulations were conducted. In each simulation, 8 datapoints (the number of participants in Schwiedrzik et al., 2009) were selected at random from two distributions of d-primes with the means and SD corresponding to the 'Pre-Training' and respectively Block 6 in Schwiedrzik et al. (2009). A difference score was calculated for each of the 8 'participants' by subtracting 'Pre-training' from 'Block 6'. These difference scores were entered into a Bayesian independent-sample two-tailed t-test with a 'wide' prior, against the d-prime difference scores in the current experiment (i.e., Day 3 minus Day 1), separately for the Learning and Control conditions. It was not possible to compare d-prime changes after 1000 trials in their experiment (i.e., 4th block of the 2nd session), because the d-prime was not reported. Consequently, an additional test was run against the difference scores between Block 6 in the current task (from Day 2) and Day 1, for the Learning group only.

The distributions of BFs obtained are illustrated in Figure 16A. Only 0.1% of analyses showed at least BF > 3 for a difference between the present experiment's data and the simulated data. Conversely, 31.6% found BF > 3 against a difference between Learning and the simulated data (13.5% compared to Control). Comparing to the increase between Day 1 and Block 6 in Day 2 yielded a similar pattern to the Control group. To explore whether the high percentage of inconclusive analyses is given by the small sample size in Schwiedrzik et al. (2009), the simulations were repeated, this time sampling the same number of included datasets as in the present experiment (12 in each condition). Distributions and percentages are illustrated in Figure 16B. Increasing the sample size in the simulated data strengthened the conclusion of no difference between the improvements in d-prime sensitivity obtained in the Learning group and those observed by Schwiedrzik et al. (2009), however it did not

substantially affect the pattern of comparisons with the Control group. Altogether, while most comparisons are inconclusive, for all comparisons the probability of finding evidence against a difference is substantially higher than finding evidence for a difference.

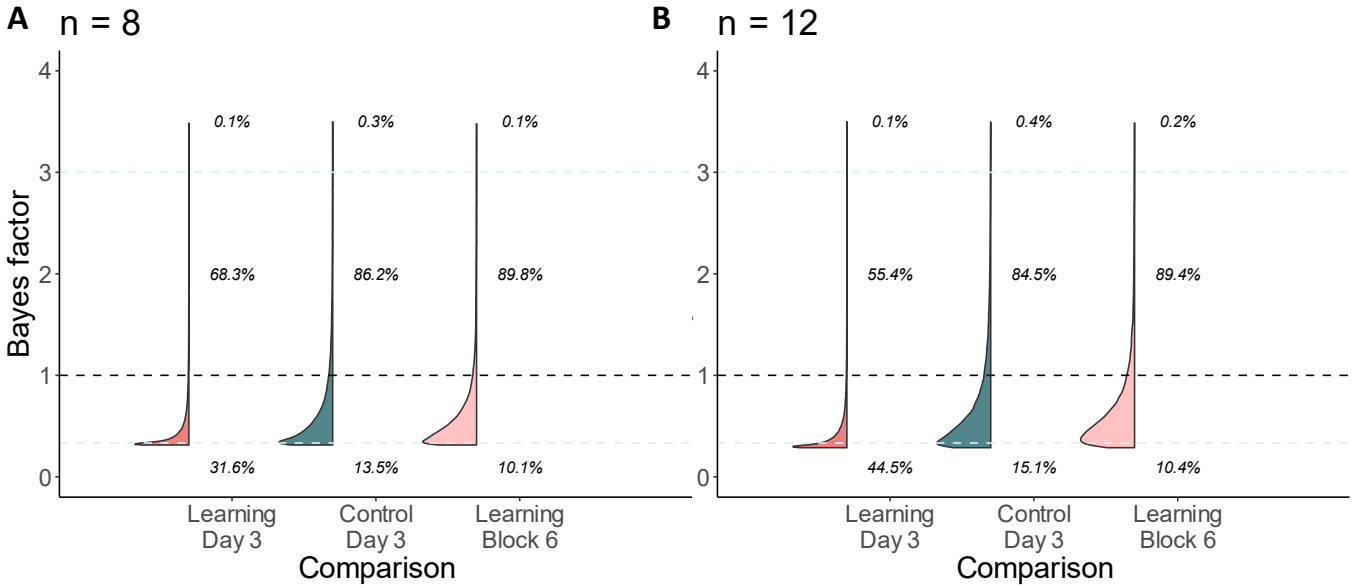


Figure 16. Distribution of BF from Bayesian independent samples t-tests between mean differences (Day 1 compared to Day 3 for Learning and Control, and Day 1 compared to block 6 in Day 2 for Learning only) in the present experiment, and simulated data points based on descriptive statistics in Schwiedrzik et al., (2009). Panel A shows comparisons to 8 simulated data points, and Panel B for an equal number of data points as the current groups (n = 12). Because less than 1% of analyses showed BF above 3, the distributions only show values until BF = 3.5. Horizontal dashed lines mark the interval of BF deemed inconclusive, between 1/3 and 3. BFs lower than 1/3 indicate moderate evidence for the null hypothesis (no difference between mean differences).

3.4.5 Q4E. Is there a change in performance within the first measurement session?

One indirect way to assess if completing the threshold measurement session in Day 1 drove the observed improvements is to test for changes in performance throughout the MoC session, for example by comparing 'windows' or segments of the task. Another way to test for Day 1 increases could also be to compare the mean stopping contrast from the MoL with the PF inflection point from the MoC. However, this latter approach has the limitation of the two indices likely not being conceptually comparable, given the differences in the methods

of obtaining them (for example, PAS answers were considered as part of the stopping rule in MoL, Appendix 5) – although there is evidence that the two thresholds can converge, under a PF model other than the Weibull model used here (Herrick, 1973). Indeed, the mean stopping contrast in MoL tended to be lower than the MoC-based PF inflection point – which would be more consistent with an explanation that the two points might index different levels of performance rather than that the thresholds increased in the span of the session.

In any case, observing such increases in performance within Day 1 would be consistent with the explanation that the learning was driven by practice during this session. However, not observing them does not necessarily challenge this explanation, because a few factors could explain a lack of increase. First, segmenting the session for analyses purposes (as the first alternative above would involve) means that the performance estimates are based on fewer trials, thus introducing the possibility of higher variability which in turn could mask small increases. Here, only 6 trials were presented at each stimulus contrast in each block of the MoC task. Secondly, it could be the case that any possible improvements are latent and hence not observable immediately after practice. This possibility was supported by empirical evidence (further discussed in section 3.5) from other visual learning tasks - such as Stickgold and colleagues who found that thresholds (defined as SOA yielding 80% performance in a line discrimination task) did not improve in re-tests conducted in the same day 3-12h later, but did improve overnight when the sessions were separated by sleep (Stickgold et al., 2000).

Given the limitation with the comparison between MoL and MoC, only the comparison across MoC blocks was further explored. The discrimination data from the MoC task in Day 1 was analysed block-by-block, using the same PF model and parameters as in Q1P. The focus on PFs rather than d-primes was justified in the context of the planned analyses specifically targeting PFs. Because there was support for the null hypothesis of no difference between the Test and Catch conditions in discrimination thresholds (section 3.3.1), the data was pooled together across conditions. A Bayesian ANOVA found very strong evidence against an effect of block number, $BF_{null} = 2.2 \times 10^2$. Planned pairwise comparisons (one-tailed paired t-tests with a 'wide' prior) targeting the early stages of learning (100-200 trials) also found strong evidence for the null, when comparing block 1 with blocks 2 ($BF_{null} = 18.9 \pm 0.07\%$) and 3 ($BF_{null} = 17.8 \pm 0.17\%$). Altogether, these results do not show support for

improvements during Day 1 – although it is important to consider, when interpreting them, the caveats discussed above in relation to latent consolidation and possible noise in the parameter estimates.

3.5 Discussion

This study, using a paradigm manipulating contrast rather than SOA, expanded upon the previous claim (Schwiedrzik et al., 2009, 2011) that repeated exposure to a stimulus individually calibrated to likely not be consciously perceived leads to an improvement in objective discrimination and clarity of subjective experience. The planned analyses replicated the results reported by Schwiedrzik and colleagues, and found strong Bayesian evidence in the Learning group that performance improved from before to after training, in both measures they collected (discrimination, Q1P, PAS, Q3P) and an additional objective measure (detection, Q2P). Crucially though, the same improvements were observed in the Control group who did not complete the training. In some cases (accuracy from the PF at the trained contrast), evidence for an increase from Day 1 to Day 3 was present in the Control group but not the Learning group – although this is possibly explained by the slightly lower number of included participants in the Learning group than in the Control group. In any case, for all comparisons, there was evidence against the hypothesis that the improvements were larger in the Learning group than the Control.

Further checks for any improvements during the training session in the Learning group found that performance improved between Day 1 and the first block of training in Day 2 but not throughout the rest of Day 2, or between Day 2 and Day 3 (Q2E). However, between-group comparisons of d-prime sensitivity at the trained contrast (Q1E) yielded mostly inconclusive results, suggesting that more data would be needed. The inconclusive d-prime but not PF-related findings might also be explained by PFs being more robust against spurious variations in sensitivity at each individual contrast level. Nevertheless, these findings have a few interesting implications. First, as Schwiedrzik and colleagues also reported, perception improves with training, with this conclusion unaffected by the choice of measurement of learning (objective vs subjective). Secondly, I deem unlikely that such improvement is due to repeated exposure to sub-threshold stimuli.

What could explain the improvements from Day 1 to Day 3, if not the training in Day 2? Two possible mutually non-exclusive factors are apparent. First, the effects could be due to practice during the Day 1 PF measurement sessions, consistent with the outcome of Q2E showing no improvement took place during the training session. This is not improbable, given the high number of trials collected here for accurate estimations of the PF (468). The

explanation is consistent with previous findings of PL where the Pre-Post measurement sessions also used a high number of trials (640 in Schwiedrzik et al., 2009, 720 in Schwiedrzik et al., 2011, 864 in Nishina et al., 2007, around 800 trials in Karni & Sagi, 1993). Conversely, no improvements were reported at the trained parameter nor the control condition (no training) when only 160 above-threshold trials were displayed (Watanabe et al., 2001). Where improvements were not observed despite using a comparably high number of visible measurement trials (720 in Seitz et al., 2005), this outcome could be due to experimental design choices; Seitz and colleagues (2005) collected two pre-training measurements sessions, with the second considered as 'Pre' despite large baseline changes between the two 'Pre' sessions being mentioned. Therefore, it is also likely that comparisons between the first 'Pre' session and the 'Post' would have showed improvements. Altogether, this evidence suggests that the improvements observed in the present study, and likely in the previous attempts to demonstrate VPL from unconscious stimuli, were driven by the extended practice during the measurement sessions before training. The exact mechanism of this effect would need to be further explored. Nevertheless, if this explanation is correct, then both unconscious and conscious trials might have contributed to the improvements in the present study and in the previous literature. This is because the MoC approach involves presenting stimuli with different visibilities, including fully visible, so only some of these trials were 'unconscious'. Therefore, none of the experimental designs could disentangle between contributions. It can be speculated though that since most of the trials pre-training tend to be at least somewhat visible, especially in MoC paradigms, the contribution of objectively-unconscious trials is likely minimal. In any case, the conundrum for unconscious learning studies might therefore be that the measurement itself might be affecting the phenomenon of interest. Unfortunately, it is not straightforward to design a psychophysical experiment with a number of trials low enough to not produce learning but high enough to produce accurate PF measurements. While other methods to estimate PF parameters such as QUEST+ (Watson, 2017) would require 64-128 trials, there is no guarantee that this number is sufficiently low. Indeed, Hussain et al. (2009) found that 105 trials of unmasked training on textures of different contrasts was sufficient to generate an improvement in a discrimination task compared to a control group who received no training.

The second possible, albeit more speculative, explanation involves the sleep between sessions. In the current paradigm, all experimental sessions occurred on consecutive days, presumably separated by sleep for all participants. The literature mostly supports an effect of sleep on VPL (reviewed by Walker & Stickgold, 2004) – for instance, effects of sleep were reported for contrast discrimination (Mednick et al., 2008), and line texture discrimination (Karni et al., 1994; Karni & Sagi, 1993; Stickgold et al., 2000), but not for noise texture discrimination (Hussain et al., 2009). Karni and Sagi (1993) further reported that while there were substantial decreases in the SOAs yielding 80% discrimination accuracy in the 24h period following training with mostly visible stimuli (600 trials at SOAs yielding 80% accuracy plus around 200 trials of accuracy under 60%), there was no improvement in the first 6-8 hours post-practice. These findings, together with others (e.g., Stickgold et al., 2000), therefore suggested a latent, sleep-dependent consolidation period characterized by fast performance gains – explanation consistent with the present findings (Q1E and Q2E) that discrimination sensitivity increased between Day 1 and the first training block in Day 2 (separated by sleep) but not within either Day 1 (Q4E) or Day 2. However, it is not fully consistent with Schwiedrzik et al.'s (2009) findings from their Day 1 (which consisted of both pre-training and the first training session, hence not separated by sleep). There, performance at the trained contrast was not different from 0 in either pre-training nor the first block of the training session, but improvements were reported within the first training session. As for why improvements did not occur in the current study between Day 2 and Day 3, which were also separated by sleep, one possible explanation is that in order to trigger latent consolidation, stimuli might need to enter awareness during training. Nevertheless, the effect of sleep on VPL is far from well-understood, and likely dependent on the specific task. Future work could aim to map further the time course of the latent consolidation for each paradigm and parameter of interest, and either model the expected improvement and adjusting the training parameters, or beginning the training after the consolidation plateau.

It is important to consider the broader implications of these findings for the consciousness literature. Since there was no added benefit of the training, the current results do not seem consistent with a key prediction from Cleeremans and colleagues' radical plasticity thesis/SOMA that learning should occur from stimuli initially not consciously experienced. Nevertheless, as discussed above, it could be the case that some learning occurred from

unconscious trials. It might also be the case that with more training, there would have been an effect – however, in theory, there is no upper limit for how long this training would need to be. This consideration challenges the falsifiability of this SOMA prediction, and adds to the difficulty in empirically supporting this theory (Cleeremans et al., 2020). On the other hand, the current results could be interpreted as more consistent with the common prediction from GWT and RPT, that learning at a behavioural level cannot occur in the absence of awareness. Even so, the two theories propose different underlying neural mechanisms for awareness (for a review of ToCs, see Seth & Bayne, 2022). GWT suggests that information breaks into awareness when local processing (e.g., within early visual areas) is sufficiently strengthened by local recurrencies to be 'broadcasted' to a global neuronal workspace (GNW) comprised by areas such as the prefrontal cortex – a process called ignition (e.g., Dehaene et al., 2006). With regards to VPL, GNWT would then predict that even if a stimulus activates local recurrent processing, it cannot induce learning without the global ignition, however it does not specify where the locus of the learning might be. Lamme's RP account (Lamme, 2006) argues that the local recurrent processing itself is sufficient for both learning and awareness, without the need of a global ignition. Because the differences are at a brain rather than behavioural level, one limitation of the current results is that they cannot be used to disentangle between these two theories. More broadly though, neither the current nor Schwiedrzik and colleagues' work was directly, a priori aimed at testing predictions from any ToCs, so apparent agreement or disagreement with any predictions does not allow strong conclusions about the viability of any ToC.

In any case, a finding that unconscious (i.e., sub-threshold) stimuli cannot drive learning could have implications for ToCs that target the neural correlates of consciousness, when considering the brain mechanisms possibly supporting contrast VPL. While these mechanisms are yet to be fully mapped, there is converging evidence from both human and animal studies that the locus of the learning is at the very early visual processing stages. For example, Hua et al. (2010) trained cats to discriminate whether a grating of a near but above-threshold contrast presented monocularly was oriented to the left or right. They found that both behavioural performance and contrast sensitivity in the primary visual cortex (V1) cells associated with the trained eye increased with training, and only partially transferred to the untrained eye. In human fMRI studies, increases in contrast detection and

discrimination performance were accompanied by increases in brain activity for the trained stimulus, as indexed indirectly by the blood oxygen-level-dependent (BOLD) signal, in V1 (Furmanski et al., 2004) and even earlier, in the lateral geniculate nucleus or LGN (Yu et al., 2016 – although no effects were reported in V1, possibly due to task differences). In any case, before investigating the possible link between the neural correlates of consciousness and those of VPL, more experimental work is needed to first understand how each process is supported by the brain.

Nevertheless, VPL remains a plausible candidate role for awareness, and provides a rich methodology for further research that can be used to shed light on why awareness might be necessary for learning.

Chapter 4

4.1 Introduction

> *"any argument for why subjective reports seem a sine qua non for consciousness research is not an argument for any subjective reporting being precise or trustworthy" (Overgaard & Sandberg, 2021, p. 2)*

When the Perceptual Awareness Scale (PAS) was first introduced by Ramsøy and Overgaard (2004), possibly universal properties of awareness were hinted at. In their experiments, which stemmed from the aim to align consciousness research methodology with the view that awareness is graded rather than all-or-nothing, participants reported the degree of clearness of experience for each feature (shape, colour, position) of basic shapes. Crucially, they were told they could classify the clarity of their experiences in how many categories they wanted, using whichever labels, and only being given as suggestion the start and end points of the range ("no image at all" to "a clear image"). All five participants converged to using a 4-level scale, with the same two additional intermediate points: "a brief glimpse", and "almost clear image", besides the suggested start and end points. Strictly referring to the contents of awareness rather than the overall levels of consciousness (e.g., wakefulness, sleep etc.), the authors argued that using dichotomous a seen/not seen task cannot sufficiently characterize awareness. This is because they failed to find evidence for higher-than-chance accuracy in identifying the stimulus shape in trials rated with "No Experience" on their PAS, but not in a "not seen" condition ("No Experience" plus "Brief glimpse" trials). The authors further argued that a dichotomous approach may lead to erroneously concluding that there is above chance performance without awareness in some tasks, thus implying the need to re-evaluate previous findings from the light of a graded awareness scale. Besides these conclusions, Ramsøy & Overgaard's (2004) findings are logically consistent with two other hypotheses: that the clarity of awareness is indeed graded rather than all-or-nothing, and that awareness can be summarized for each and all individuals in the same broad degrees of clarity – although the number of distinct degrees might differ across samples and stimuli (Overgaard & Sandberg, 2021). These ideas, if supported, would have substantial implications for both theoretical and empirical consciousness research. For example, the hypothesis of awareness being graded can be seen as inconsistent with the Global Neuronal Workspace Theory (Dehaene et al., 2006). It is also inconsistent with

evidence that it is all-or-nothing, such as the findings from Sergent and Dehaene (2004) that during an attentional blink task participants' subjective-visibility answers clustered around the start and end of a continuous scale regardless of the visibility manipulations. Del Cul and colleagues (2007) found similar results in a masking paradigm. Conversely, the existence of discriminable universal awareness levels would further motivate the pursuit of new research questions, like comparing brain activity associated with different levels of awareness (e.g., Eiserbeck et al., 2022).

The PAS has now become a popular measure in consciousness research, with the original paper amassing over 300 citations (Springer article page, 26/11/2023) and prompting direct recommendations in the literature to use the scale (Prochazkova et al., 2022). This is despite there being, to my knowledge, no attempts at replicating the scale validation on a larger sample, even though in the original research (Ramsøy & Overgaard, 2004), only 3 out of 5 participants tested came up with the four clarity levels. Two participants started with additional ratings that were dropped because of underspecified definitions, and Overgaard et al. (2004) also reported that two out of seven participants were fully excluded for inconsistent use of their own categories. Even if it is assumed that the four-level classification of clarity generalizes to a large sample, and before considering further implications that this classification might have, a more fundamental question is how well the PAS relates to existing measures that have been used as indices of awareness. Two criteria are frequently mentioned when assessing PAS, both in relation to objective measures, as in those where a correct answer exists, such as detection or discrimination (Andersen et al., 2019; Sandberg et al., 2010; Szczepanowski et al., 2013; Wierzchoń et al., 2014). The first is how well changes in objective measures correlate with changes in subjective experience, hereafter referred to as 'sensitivity'. The second is whether trials identified as 'No Experience' show chance performance, hereafter referred to as 'exhaustiveness'. Andersen and colleagues (2019, p. 60) summarize these two criteria in relation to the PAS:

> *For a scale to be exhaustive, the scale must provide evidence that when participants claim to have no experience and no knowledge about what was shown (Table 1: No Experience), their performance should not be different from chance-level performance. For a scale to be sensitive, the scale must provide points such that when participants claim to have some degree of experience and knowledge (Table 1:*

Weak Glimpse, Almost Clear Experience and Clear Experience), their performance should correlate with the clarity of the experience and amount of knowledge reported.

Finding evidence for the 'exhaustiveness' or 'sensitivity' of PAS from its relationships to performance is often interpreted as evidence for the fitness of the PAS as a valid measure of awareness. For example, using the same operational definition of exhaustiveness (i.e., chance performance at PAS1), Sandberg and colleagues extended its interpretation: '*it is unclear which measure is most exhaustive, that is, which method reveals the most conscious processing'* (p. 1071). However, doing so places the onus on objective measures to be the benchmark or 'gold standard' for assessing awareness, and suggests they are intrinsically more reliable or less prone to bias than subjective measures. This assumption is rooted in complex theoretical and methodological considerations, which will be returned to in the discussion (4.4). Nevertheless, one can assess the previous evidence (i.e., analyses results) associated with exhaustiveness (chance performance at PAS1) and sensitivity (correlations between PAS and performance) without implying agreement with the assumption of task performance as the benchmark, or with the interpretation of the two labels in relation to validity. This is simply because both the PAS and task performances are claimed to test the same concept (awareness); an equally reasonable approach would be to systematically compare PAS with other trial-by-trial subjective measures like CR, although this approach would still require asking participants to make an objective judgment. Therefore, for the rest of the chapter, the two labels will be used strictly for distinguishing the two lines of analysis.

Under this interpretation, the evidence seems to support PAS sensitivity. One early study (Sandberg et al., 2010) found that PAS correlated better with performance than other subjective measures described in Chapter 1, such as confidence ratings (CR) and post-decisional wagering (PDW) in a backward masking experiment. Wierzchoń et al. (2014) also reported that PAS was better correlated with verbal identification of backward-masked faces than CR, PDW, and feeling-of-warmth (FOW), although this advantage was only present in very narrow circumstances (trials with stimulus durations over 48ms and with the objective measure presented before the subjective measure). Siedlecka and colleagues (2020) also found that variations in PAS were closely following variations in objective accuracy. However, PAS might prove the most sensitive measure only in certain circumstances; for instance, for

stimuli such as fearful faces, CR rather than PAS showed the best correlation with performance (Szczepanowski et al., 2013). While this is a promising start, currently there are no large-scale assessments of PAS sensitivity across experimental paradigms, which is arguably necessary for better understanding what this measure targets.

The evidence for above-chance performance at PAS1 is more mixed. Some features, like stimulus position (e.g., all six experiments in Jimenez et al., 2019; Overgaard et al., 2004), were found to result in above-chance performance in PAS1 trials. The same was found for shape discrimination by Sandberg et al. (2010) when pooled across all stimulus durations, albeit the difference with chance level was small, and accuracy at each stimulus duration varied substantially and non-linearly. Yet, they concluded that PAS was the most exhaustive of the three measures tested since the other measures yielded higher accuracies at their lowest levels. Other studies also reported no significant differences between performance in trials rated with PAS1 and chance, for different features of the stimuli such as shape (Overgaard et al., 2004; Ramsøy & Overgaard, 2004), colour (Overgaard et al., 2004), or whether the target digit was odd or even (Andersen et al., 2019). The caveat of the examples above though is that they used frequentist statistics, thus conflating absence of evidence with evidence of absence. In typical frequentist analyses, a p-value under 0.05 alone does not allow concluding that there is no evidence of a difference. To achieve this conclusion, different approaches like Bayesian analyses are required – an argument discussed in detail by Dienes (2015). It is therefore not yet known if the previous evidence that PAS is exhaustive allows drawing this conclusion. Moreover, the divergent findings from different experimental designs warrant further investigation of which circumstances, if any, PAS does not show exhaustiveness in. Another way to test if participants' PAS1 ("No experience") ratings indicate that they are not able to make correct judgments about the stimulus more often than chance is arguably to investigate cases where chance performance is guaranteed, namely in catch trials where nothing is presented. It would therefore be expected that almost all their answers on the PAS would be 1, with some small variation due to lapses or incorrect button presses. However, some initial evidence from previous papers (e.g., Jimenez et al., 2018) reported that more than 20% of trials had ratings higher than 1. Before attempting to infer what could have caused such answers, it would be informative to assess if this pattern is present consistently and to what extent it might vary across samples.

Nevertheless, even if PAS is found to have exhaustiveness and sensitivity, what would be the added benefit of using it over the objective measures it is compared against? Without clear advantages over objective measures, or a way to assess the inherent bias present in any behavioural response (unlike, for example, d-prime in signal detection theory) it was proposed that *"the PAS is neither new, exhaustive, unbiased, nor natural"* (Irvine, 2012, p. 645). While previous studies that conduct such comparisons do not explicitly address this question, one limitation of objective measures that the PAS addresses is that they cannot be used to assess awareness trial-by-trial, and can therefore only be used for conclusions at the condition level. An experimental design employing objective awareness measures must therefore a priori define conditions expected to result, for each participant, in null sensitivity (or chance accuracy). This approach is not compatible with all research questions nor an easy feat, since the optimal stimulus parameters for rendering stimuli at null sensitivity would differ with task, stimuli, task difficulty, consciousness manipulation methods, participants (Albrecht et al., 2010), and practice with a task (Chapter 3 of this thesis, Schwiedrzik et al., 2009, 2011). Having a reliable trial-by-trial measure of awareness should in principle benefit the field by helping narrow down only cases (i.e., trials) where awareness is absent, and also allowing the study of new research avenues, such as the fluctuations in conscious experience and brain activity when the stimulus parameters are held constant (e.g., the liminal prime paradigm, Lamy et al., 2015). Therefore, it seems justified to assess if this measure can be the PAS or not.

The current work draws on data from tasks where both PAS and objective measures were collected trial-by-trial on the same stimuli. The aim was to test if previous findings about PAS meeting the criteria for exhaustiveness and sensitivity, based on comparisons with objective measures, are robust across experiments with different experimental designs and stimuli. This approach allowed capitalizing on the availability of data from 'unconscious perception' paradigms, which aim to test a dissociation between a trial-by-trial measure of awareness (like the PAS or a binary seen/not seen task) and a measure of processing. 23 experiments were identified, across 15 studies (including Exposure Stage data from Experiment 2 in Chapter 2 and data from Chapter 3 in this thesis), with data either publicly available or shared by the authors. This allowed addressing the following questions, each corresponding to a planned analysis:

Exhaustiveness

1. Do trials that fit a definition of 'subjective unawareness' based on the PAS (i.e., PAS = 1 or "No Experience") also show chance objective performance?
2. What experimental design choices, if any, influence exhaustiveness?
3. When performance can only be at chance, in stimulus-absent catch trials, what proportion of trials are rated as 'subjectively unaware'?

Sensitivity

4. How strongly do PAS ratings predict accuracy?
5. Is the change in performance across PAS ratings gradual, or all-or-nothing?

Below is a brief overview of the experimental design and results of each of the studies included (see Table 10 for a comparative summary), except data from this thesis for which full descriptions of methods and findings can be found in earlier chapters. Only relevant experiments, conditions, and findings linked to the current analyses will be referred to.

Andersen et al. (2019) investigated the effect of expectations on objective performance and subjective experience of masked numbers. On each trial, participants saw a suprathreshold cue consisting of 2, 4, or 8 digits simultaneously presented, one of which was the target digit later in the trial. After a delay of 1000ms, the target digit was presented for a variable duration, and was immediately followed by a noise mask consisting of random lines. Participants indicated if the masked target digit was even or odd, and completed a 4-level PAS. They concluded that the PAS was exhaustive because the accuracy in PAS1 trials was around chance, for all but the longest target duration (70.6ms where it was around 70%) - although no statistical comparisons to chance were reported.

Derda et al. (2019) presented participants with backward-masked digits with a duration of 50ms and coloured in either red or blue. On each trial, participants completed either a low-level task (i.e., report the colour of the digit) or a high-level task (i.e., report if the digit is higher or lower than 5), and then a 4-level PAS. They found that across all regression models tested, PAS was a significant predictor of accuracy (reported in Supplementary Materials S2), and also that accuracy at PAS1 was higher in the high-level condition compared to low-level.

Jimenez et al. (2018) investigated the linearity of the relationship between PAS and accuracy in two tasks requiring different levels of processing. In each trial, participants saw one of the 4 possible stimuli (6, 9, I, H) in one of the corners of the screen. Then, a mask was presented for a variable duration, followed by a 4-alternative forced choice (4AFC) task to either choose which corner the stimulus appeared in (low-level) or the identity of the stimulus (high-level), and finally a 4-level PAS. They found that mean accuracy increased linearly with PAS ratings for both detection and identification, supporting earlier claims of linearity. Although no statistical comparisons to chance were reported, mean accuracy for PAS1 trials was reported to be exactly at or slightly under chance level for the high-level task, but higher than chance for the low-level task.

Jimenez et al. (2019) studied, across 6 experiments, the effects of varying stimulus duration and task type on PAS ratings and accuracy. In half of the experiments, participants responded with the location of a backward-masked square located in one of the four corners of the screen, as well as a 4-level PAS. In the other half, participants saw the letter I or H, again in one of the four corners, and had to respond with the location, PAS, and letter identity. They found that letter identification was not significantly different from chance in trials rated with PAS1. However, this occurred only for the lowest two stimulus durations, suggesting an interaction between stimulus duration and level of processing, and the conclusion of no difference was based on frequentist statistics. On the other hand, location detection accuracy in PAS1 trials in all experiments was above chance.

Jimenez et al. (2021) showed participants line drawings of animals and objects presented for variable durations, followed by a line mask for variable durations. Then, participants responded with either the colour (red/blue, low level task) or the category (animal/object, high-level task), and a 4-level PAS. They reported evidence for a linear increase in accuracy across PAS ratings in both the low- and high-level task, thus suggesting a close mapping between the two measures.

Jimenez et al., (2023) tested, across three masked priming tasks, whether global shapes can be processed in the absence of awareness. The primes were Navon figures, in which a ‘local’ shape (e.g., triangle) is repeated and arranged in a way that generates a ‘global’ shape (e.g., a square). If spatial integration and global shape processing are possible in the absence of awareness, then briefly flashing a Navon figure followed by a mask should shorten response

times to a subsequent target shape when the target shape is the same as the global prime shape. The effectiveness of the mask was manipulated by changing the ISI and mask durations. Combinations of ISI and mask durations were tested between-subjects in different experiments, each with three different tasks: single task block (2AFC on target shape), dual task block (2AFC on target shape, 4-level PAS for prime), and a prime visibility block where only the masked primes were presented, followed by a 2AFC on prime shape and 4-level PAS. Only the prime visibility block at each SOA was included in the current analysis.

In Sand and Nilsson's (2017) experiment, participants viewed a prime word ('red' or 'blue') for 6ms, followed by a blank screen, a mask, and a coloured target square. The prime could either be congruent (i.e., the word 'red' predicting a red square target) or incongruent with the target. Participants reported the colour of the target, and then completed a 'prime assessment grid', where they picked one of the six squares corresponding to a combination of the objective answers (red/blue) and a 3-level PAS. Accuracy in PAS1 trials was reported as 52%, and Bayesian evidence weakly supported the null hypothesis that sensitivity (d-prime) in PAS1 trials was not different from 0.

Siedlecka et al. (2020) investigated whether providing feedback on an objective task influenced subjective awareness. Participants saw a centrally presented Gabor patch, in a contrast calibrated individually to yield 70% accuracy. Then, they were asked to choose if the patch was left or right-oriented, and to complete a 4-level PAS. Feedback for the left-right answer was provided only in the Feedback condition. They reported that accuracy for trials rated with PAS1 was higher than chance, and that PAS predicted accuracy similarly across both feedback conditions.

Skewes et al. (2021) investigated whether providing incorrect feedback changed participants' accuracy and subjective experience in a masking task. The present analyses included only the True Feedback condition. Each trial presented two Gabor patches, one with an orientation of 0 (i.e., vertical) and one with an orientation between 0-45deg, chosen for each participant to be discriminable on 65% of trials. Participants reported which side of the fixation cross (left or right) the tilted Gabor appeared on, and then completed a 4-level PAS and confidence ratings (CR). The order of the PAS and CR questions was randomised.

Stein and Peelen (2021) conducted a series of studies using upright and inverted faces, and different objective and subjective measurements of awareness, as well as methods of manipulating awareness. In Experiment 2 (data from the backward masking with discreet 4-level PAS paradigm only), participants saw an upright or inverted face displayed for a variable duration to the left or right of a central fixation cross, followed by 3 noise masks. They then reported the location of the face and completed the PAS. In Experiment 3, the stimuli and masks were identical to Experiment 2, but they varied the display time and interstimulus interval (ISI) between the faces and masks. Each trial was followed by three questions: the PAS, a left-right localization task, and an upright-inverted discrimination task, with the order of the questions pseudo-counterbalanced. PAS data was dichotomized and transformed in detection d-prime. The current analyses included only the upright-inverted discrimination accuracy.

Thiruvasagam and Srinivasan (2021) focused on whether the gradedness of PAS answers changes if the task requires attending to global or local features. Participants viewed a centrally presented masked Navon figure for variable durations between masks of 250ms each. Navon figures consist of a global shape (e.g., the letter S) composed from smaller shapes (e.g., the letter H). After every trial, participants were asked to report the letter they saw, either at global or local level depending on the block, and they completed a 4-level PAS.

Table 10. Summary overview of the included experiments, referring to included participants, conditions, and measurements only.

Study	Measurements	Stimuli	Stimulus duration(s)	Mask duration(s)	Awareness manipulation	Feedback	Chance level	Task type	N participants
Andersen et al., 2019	2AFC digit as odd/even, 4-level PAS	Single digits	11.8ms, 23.5ms, 35.3ms, 47.1ms, 58.8ms, 70.6ms	352.9ms	Backward masking	No	0.5	High level	n = 58 (Exp 1: 29, Exp 2: 29)
Derda et al., 2019	2AFC digit higher/lower than 5, or colour (red/blue), 4-level PAS	Single digits	50ms	250ms	Backward masking	No	0.5	Both	n = 41
Jimenez et al., 2018	4AFC location detection or identification of letter or number stimuli, 4-level PAS	Letters or digits	13.3ms, 26.7ms, 40ms, 106.7ms	667ms, 680ms, 693ms	Backward masking	No	0.25	Both	n = 16
Jimenez et al., 2019 – Study 1 (Exp 1-3)	Exp 1-6: 4AFC location, 4-level PAS	Squares	13.3ms, 26.7ms, 40ms	26.7ms	Backward masking	No	0.25	Low-level	n = 24 (Exp 1 and 5), n = 23 (Exp 2 and 4), n = 21 (Exp 3 and 6)
Jimenez et al., 2019 – Study 2 (Exp 4-6)	Exp 4-6 only: 2AFC of letter identity (I/H)	Letters	26.7ms, 53.3ms, 80ms				0.5	High-level	
Jimenez et al., 2021	2AFC of colour (red/blue) or category (animal/object) of line drawings, 4-level PAS	Line drawings	13.3ms, 26.7ms, 40ms, 53.3ms, 66.7ms	640ms, 666.7ms, 653.3ms, 680ms, 693.3	Backward masking	No	0.5	Both	n = 26
Sand & Nilsson, 2007	2AFC prime word identification ("red"/"blue"), 3-level PAS	Words	6.25ms	143.75ms	Backward masking	No	0.5	High level	n = 66
Siedlecka et al., 2020	2AFC direction of grating (left/right), 4-level PAS	Gratings	33.3ms	No mask	Short stimuli	1 condition	0.5	Low level	n = 37
Skewes et al., 2021	2AFC side of the screen (left/right) that showed a tilted grating, 4-level PAS	Gratings	33.3ms	600ms	Backward masking	Yes	0.5	Low level	n = 31, True Feedback condition
Stein & Peelen, 2021 – Exp 2	2AFC of face as inverted/upright, 4-level PAS	Faces	10ms, 20ms, 30ms	300ms (3 masks x 100ms)	Backward masking	No	0.5	High level	n = 24, Masking and discrete PAS condition
Stein & Peelen, 2021 - Exp 3	2AFC of face as inverted/upright, 4-level PAS	Faces	8.33ms, 8.33ms + 8.33ms ISI,	300ms (3 masks x 100ms)	Backward masking	No	0.5	High level	n = 94

			16.7ms, 25ms, 33.3ms + 8.33ms ISI						
Thiruvasagam & Srinivasan, 2021	2AFC of letter identity, 4-level PAS	Letters	16.7ms, 33.3ms, 50ms, 66.7ms 83.3ms, 100ms, 116.7ms,	250ms	Sandwich masking	No	0.5	High level	n = 22 (after exclusions)
Halchin - Chapter 2	5AFC natural scene recognition, 4-level PAS	Images	16.7ms, 1000ms	2000ms	Backward masking	No	0.2	High level	n = 48 (after exclusions)
Halchin - Chapter 3	2AFC of arrow direction (left-right), 4-level PAS	Pair of arrows	16.7ms, 16ms + 500ms ISI (in catch trials)	200ms	Metacontrast masking	1 session, blockwise	0.5	Low level	n = 35 (after exclusions)
Jimenez et al., 2023	2AFC square or diamond, 4-level PAS	Simple shapes	40ms	Forward mask: 100ms; Exp 1-3 backward mask: 40ms, 53ms, 67ms	Sandwich masking (Exp 1-3); Forward masking (Exp 4)	No	0.5	Low level	n = 80 (20 each)

4.2 Methods

4.2.1 Inclusion criteria and statistical choices

The inclusion criteria for the papers were:

1. Presence of a discrete measure of perceptual clarity (e.g., PAS, as opposed to confidence, wagering etc.) and at least one discrete objective measurement, collected trial-by-trial for the same stimuli.
2. The objective measure not having an 'I don't know' option and having a true correct answer (i.e., not a perceptual-based judgment), to allow comparison to chance performance.

For studies where more than one objective measurement was collected in each trial (detection/discrimination in Days 1 and 3 in Chapter 3, localization/identification in Experiment 3 in Stein and Peelen, 2021, localization/identification in Study 2 in Jimenez et al., 2019), only the discrimination/identification measure was included. This was done to avoid having multiple datapoints in the sample coming from the same trial (and thus being non-independent), and discrimination was chosen for all studies for consistency.

After exclusions, 327647 trials from 646 participants across 23 experiments and 15 studies remained, that were included in the analyses. Each dataset was pre-processed using custom R code, to achieve a similar format. All data pre-processing and analyses were conducted in R v4.2.1 (R Core Team, 2021) and RStudio v2021.09.1.372 (RStudio Team, 2021). The following packages were used: tidyverse v2.0.0 (Wickham et al., 2019), here v1.0.1 (Müller & Bryan, 2020), gghalves v0.1.4 (Tiedemann, 2022), readxl v1.4.2 (Wickham & Bryan, 2023), R.matlab v3.7.0 (Bengtsson, 2022b), scales v1.2.0 (Wickham et al., 2023), and BayesFactor v.0.9.12-4.4 (Morey et al., 2022). For one dataset, additional pre-processing was conducted using custom-written MATLAB code in MATLAB R2021a (*MATLAB*, 2021). All analyses were two-tailed, and assumed the prior odds to be the same for the null and alternative hypothesis, and therefore the BF is the same as the posterior odds. The prior for the effect size was set to 'medium' across all tests: for the Bayesian t-tests, this means $\frac{\sqrt{2}}{2}$ (r = 0.707), while for the general linear models, this means $\frac{\sqrt{2}}{4}$ (r = 0.354).

Where necessary, stimulus-absent catch trials (but not visible catch trials) or control conditions were removed. Stimulus-absent catch trials were analysed separately. To better understand what the columns coded for, the analysis files and additional explanatory documents (where available) were used, and/or correspondence with the authors. Generally, the same exclusion criteria that each author reported in the papers were applied. However, the criteria were changed where it was considered reasonable to do so– for example where participants were excluded from the original analyses because of issues with brain imaging data but the behavioural data was unaffected, or where issues were discovered with the authors' exclusion process. For Chapter 2 Experiment 2, the same exclusion criteria applied as in the main analyses (all based on answers in the visible/blank attention checks). For Chapter 3, only participants who failed attention checks in Day 2 were removed (based on EC2 and 3, section 3.3.1), as there were no clearly visible/clearly absent attention checks in Days 1 and 3. Links to the open data, where available, can be found in Appendix 6. BF errors were under 0.01% unless otherwise specified. Stimulus Duration and Mask Duration were re-calculated, where necessary, as multiples of the reported monitor rate, and were treated as continuous.

4.2.2 Data labelling procedure

Tasks were labelled as low-level if they relied on the stimulus appearance (e.g., colour, grating orientation etc.), and high-level if they relied on processing stimulus category or identity (e.g., whether a target digit was above or below 5, even or odd, etc.). This is in line with the conceptualization of low vs high-level tasks in the included studies that investigated both (Derda et al., 2019; Jimenez et al., 2018, 2019, 2021).

4.3 Analyses and results

4.3.1 Q1. Subjective vs objective 'unconscious'

To assess the exhaustiveness of PAS, as in whether subjectively 'unconscious' trials (rated with PAS1) had chance accuracy, for each study the mean accuracy was computed for each participant. Means were then compared to the corresponding chance level using a Bayesian equivalent of a one-sample, two-tailed t-test. Participants whose means were based on

fewer than 10 trials (i.e., where the participant had fewer than 10 trials rated with PAS1) were excluded. Participants whose accuracy averages were 3 SDs above or below the mean were excluded (n = 7, all were 3 SDs above). For tasks where the chance level differed from 0.5 (e.g., 25% in 4AFC), each mean was rescaled and centred on 0.5 – which had no effect on BFs.

There was moderate evidence against a difference between chance level and objective performance in 'subjectively unconscious' trials in only 3 out of 15 studies (Table 11 and Figure 17). In 7 out of 15 studies, there was at least moderate evidence for a difference, suggesting that performance was higher than chance. 5 studies fell within the weak to inconclusive area (BFs < 3). Including the removed outliers did not influence the pattern of results for 13 out of 15 studies. In the remaining two, the BF were initially inconclusive, but after including the outliers they indicated moderate evidence for a difference. Taken together, the analysis highlights that trials labelled with PAS1, or 'subjectively unconscious', often do not show chance objective performance.

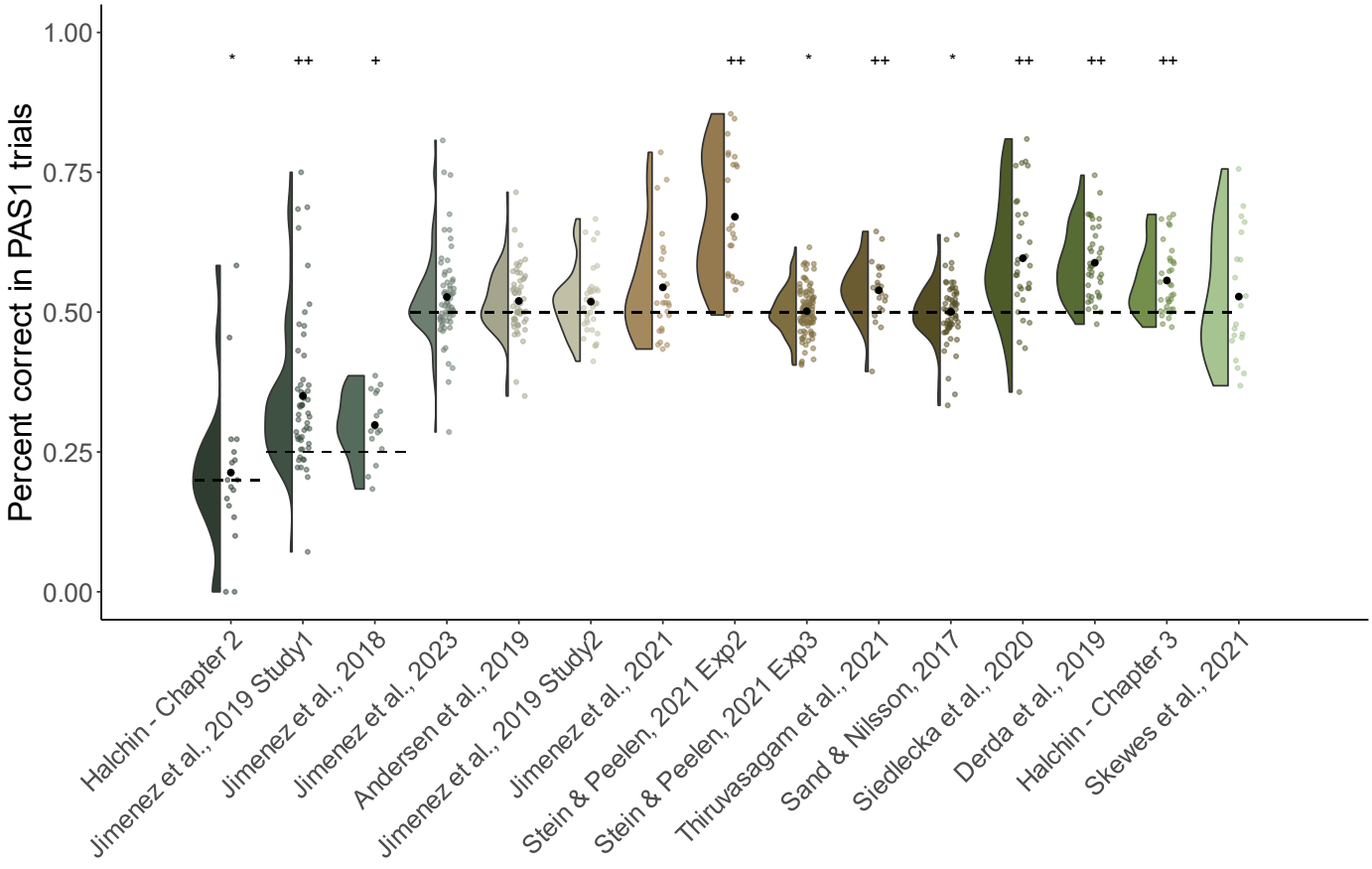


*Figure 17. Distribution of accuracy values in trials rated with PAS1 ("No Experience") in each study, after exclusions. The dotted lines mark chance level in each task (not rescaled). The black dots represent the mean of means. The following labelling convention was used: * = moderate evidence for the null (BF = 3-10), ** = strong evidence for the null (BF > 10), + = moderate evidence for the alternative, ++ = strong evidence for the alternative. No label means an inconclusive BF. Exact values are mentioned in Table 11.*

4.3.2 Q2. Factors influencing accuracy in trials rated subjectively unconscious

The results from Q1 suggest that there is substantial variability across studies in performance in PAS1 trials, and only occasionally is performance at chance (i.e., exhaustive). As discussed in the introduction (4.1), divergent findings about performance in PAS1 trials are also present in the broader literature. What experimental design factors, if any, might influence performance in these trials? To address this question, means for each included participant in each combination of factors were computed: Stimulus Duration, Mask Duration, and Task Type (high-level or low-level). Means based on fewer than 10 trials per participant per factor combination were excluded. Factor combinations represented by only one participant were also excluded. The masking method, whether feedback was provided, and the interstimulus interval (ISI) were not modelled, because in all cases, 75% or more of data belonged to one of the possible categories (backward masking, no ISI, respectively no feedback) – thus data was pooled across different levels of these factors. For studies that included sandwich masking, only the duration of the backward mask was included, for consistency with the rest of the papers. The analysis was split by Task Type to avoid non-independent data points from studies where both the low and high-level tasks were conducted within-participants, and also to avoid testing for a three-way interaction, which would be difficult to interpret. Nevertheless, some non-independence remains, given that in some studies each participant could have had data at more than one stimulus/mask duration. Appendix 6 shows that when these studies are removed and only independent datapoints remain, the regression lines and BFs are consistent with the models based on the full dataset in Figure 18.

For each Task Type, Stimulus and Mask Duration were entered in a Bayesian general linear model. Figure 18 shows the mean accuracies for each task type and predictor. For High-Level tasks, there was strong evidence for a main effect of Stimulus Duration, $BF_{alt} = 3.81 \times 10^4$, moderate evidence against a main effect of Mask Duration, $BF_{null} = 8.52$, and inconclusive evidence against an interaction, $BF_{null} = 1.95$. The model including only Stimulus Duration was the strongest, and it was preferred over the full model by a factor of 37.1. The main effect was followed up with a Bayesian equivalent of R^2 in Stan as proposed by Gelman et al. (2019). This test found that Stimulus Duration for High-level tasks accounted for 3.42% of the variance in accuracy at PAS1, as indicated by the mean of the posterior distribution of R^2 values (SD 1.21%). This was consistent with the adjusted R^2 of a linear regression (3.19%). The full model explained 3.72% of the variance (SD 1.27%, frequentist R^2 3.085%).

For Low-Level tasks, there was moderate evidence against a main effect of Stimulus Duration ($BF_{null} = 6.59$) and interaction ($BF_{null} = 3.63$), but inconclusive evidence against a main effect of Mask Duration ($BF_{null} = 2.32$). The evidence against the full model was $BF_{null} = 46.8$. Altogether, these findings suggest there is little influence of timing parameters on the accuracy of PAS1 trials in Low-Level tasks.

4.3.3 Q3. PAS answer distribution in stimulus-absent catch trials

8 out of 15 studies included stimulus-absent catch trials. Frequency distributions of answers are shown in Figure 19. An additional visualization with the percentages of trials answered with each PAS rating by each participant can be found in Figure S2, Appendix 6. On average, participants responded with PAS1 in 84.9% (SD = 9.33%) of trials, with the second most frequent answers being PAS2. Mean PAS in each study varied between 1.06 and 1.39. Across all studies, participants answered with PAS1 in at least 70%, and at most 97.2%, of stimulus-absent trials. These results suggest that while most stimulus-absent trials are labelled as expected with 'No Experience', on average the figure does not approach 100%. The highest proportion of PAS1 answers, registered in data from Chapter 2, cannot be explained by the strict exclusion criteria – while 4 out of the 12 excluded participants met the exclusion criterion of too many answers of PAS above 1 in blank trials, 2 also met other criteria which warranted exclusion anyway. Including the remaining 2 participants would have added only 6 PAS > 1 trials to the distribution, thus having minimal impact. The consistency of this

pattern across studies with different tasks and paradigms highlights that the relatively high frequency of PAS2+ answers is not an idiosyncrasy of particular studies.

*Table 11. Accuracy in PAS1 trials and Bayesian R^2 variance explained for each study. + and blue text = moderate evidence for the alternative, ++ and blue text = strong evidence for the alternative. * and orange text = moderate evidence for the null, ** and blue text = strong evidence for the null. Text with colour only and no label indicate weak BFs. No label and black text marks inconclusive BFs. The error percentages were under 0.2% for all models.*

Study	Accuracy in PAS1 trials			Bayesian R^2- PAS and accuracy
	Mean accuracy (SD)	*Direction*	*BF*	
Derda et al., 2019	0.588 (0.064)	BF_{alt}	2.9×10^{7} ++	58.9%
Stein & Peelen, 2021- Experiment 2	0.67 (0.111)		1.25×10^{5} ++	54%
Jimenez et al., 2019- Study 1	0.35 (0.133)		1.25×10^{4} ++	40.4%
Halchin – Chapter 3	0.557 (0.059)		7.78×10^{3} ++	64.1%
Siedlecka et al., 2020	0.596 (0.109)		1.10×10^{3} ++	38.9%
Thiruvasagam & Srinivasan, 2021	0.539 (0.054)		14.6 ++	79.1%
Jimenez et al., 2018	0.298 (0.061)		8.19 +	84.1%
Jimenez et al., 2023	0.527 (0.083)		2.99	41.1%
Andersen et al., 2019	0.52 (0.059)		2.97	75.9%
Jimenez et al., 2021	0.544 (0.094)		2.28	64%
Jimenez et al., 2019- Study 2	0.519 (0.060)	BF_{null}	1.15	39.4%
Skewes et al., 2021	0.528 (0.113)		2.52	60.4%
Halchin – Chapter 2	0.213 (0.142)		3.77 *	77.1%
Sand & Nilsson, 2007	0.50 (0.054)		7.18 *	79.4%
Stein & Peelen, 2021- Experiment 3	0.502 (0.042)		8.06 *	74.4%

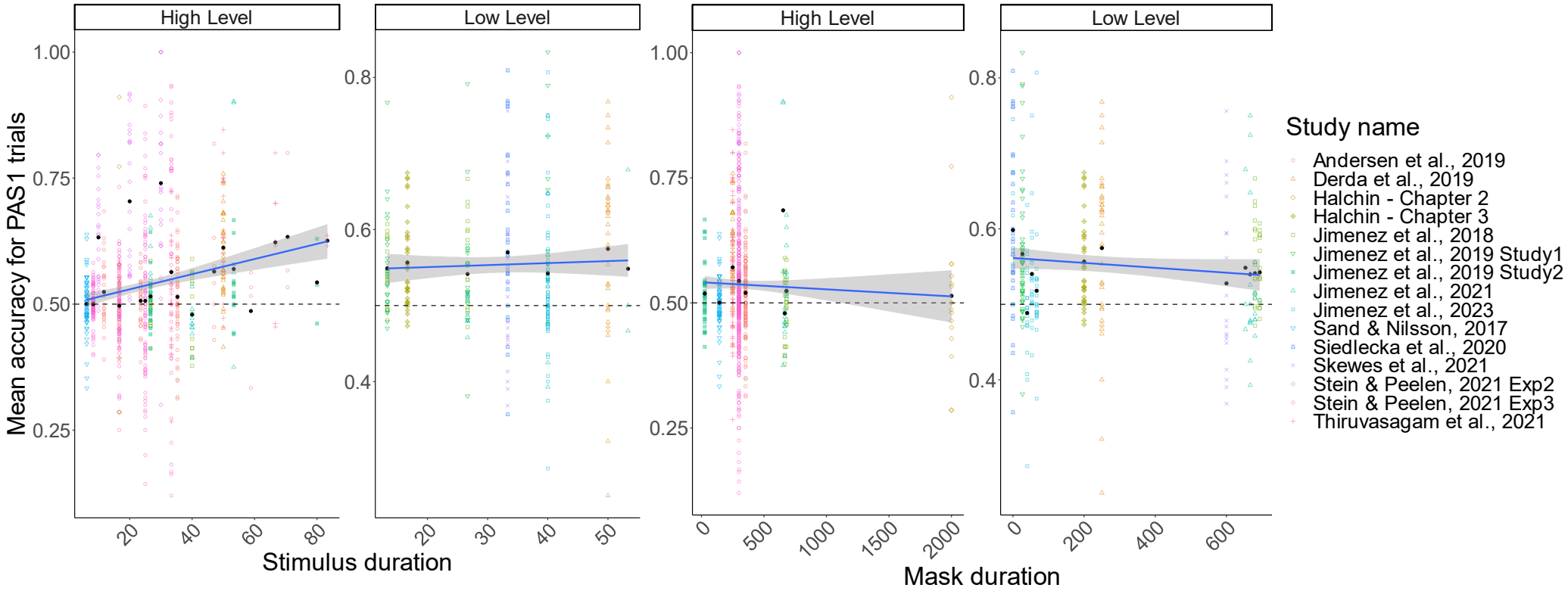


Figure 18. Distribution of accuracy values in trials rated with PAS1 ("No Experience"), for each trial type and predictor, and the corresponding regression lines. Each black dot represents the mean for the specific contrast level. The ribbon around each line shows the 95% confidence interval. The dotted line marks the chance level.

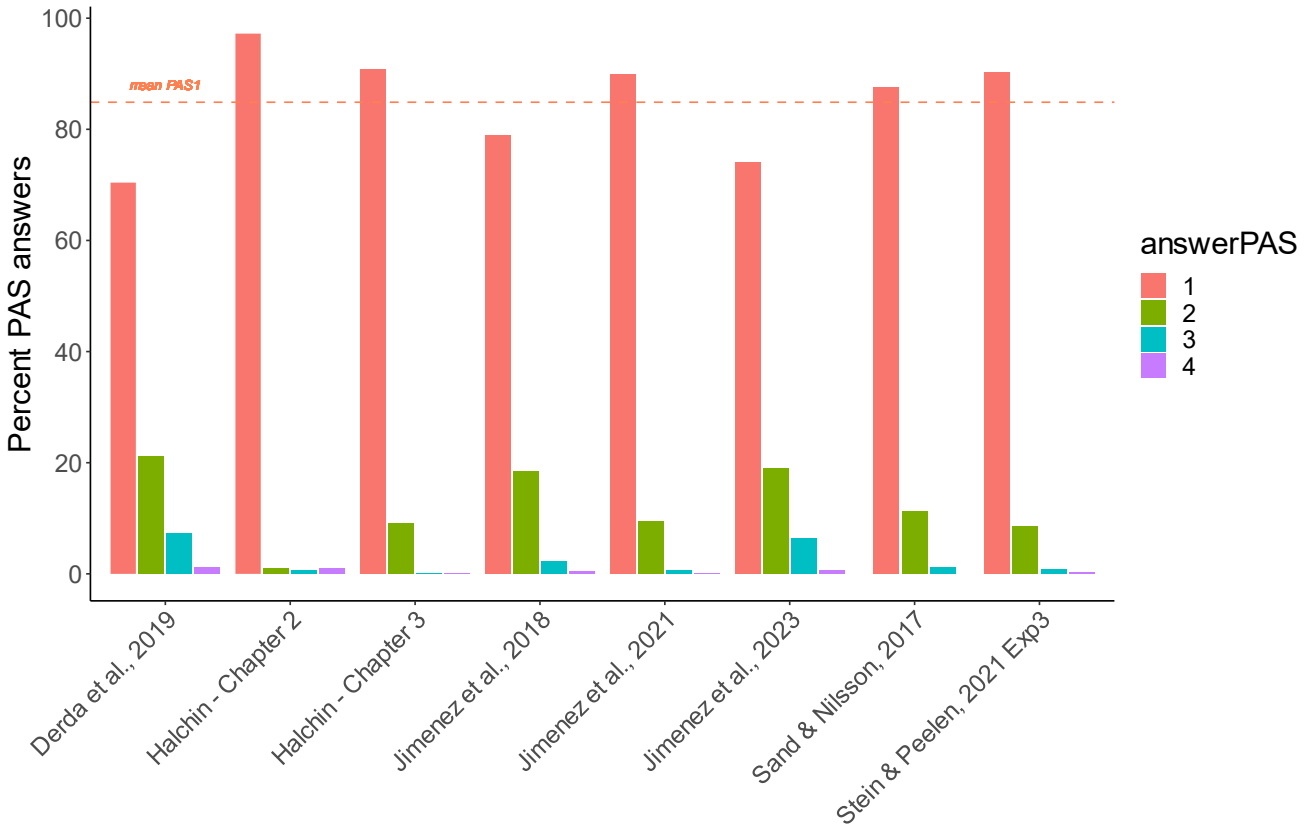


Figure 19. Relative frequency of PAS answers at each level, for each study, in stimulus-absent trials. The dashed line indicates the average across all studies.

4.3.4 Q4. How strongly do PAS ratings predict accuracy?

Next, the relationship between overall PAS ratings and accuracy was assessed. As previously discussed, if PAS met the criteria for sensitivity, then it would be expected to observe that increases in PAS are strongly linked with increases in accuracy at each level. For this purpose, a Bayesian general linear model with numeric PAS ratings and Study as fixed effects (prior = "medium") was computed. Again, means based on fewer than 10 trials per participant at each PAS rating were removed. Figure 20 illustrates mean accuracy at each PAS level, for each participant in each study. For consistency, only studies with a 4-level PAS were included. For tasks where the chance level differed from 0.5 (e.g., 25% in 4AFC), each mean was rescaled and centred on 0.5. For the purpose of the analysis and to increase the comparability of the results with other studies, the model was assumed to be linear and no logarithmic transformations were performed. Visual inspection of the data (Figure 20) also suggests that at the group level many datasets showed linear patterns. These choices are in

line with previous literature assuming or testing for linearity (Jimenez et al., 2018, 2019, 2021; Schwiedrzik et al., 2009, 2011). Where non-linear models were used in previous literature (e.g., Sandberg et al., 2011; Thiruvasagam & Srinivasan, 2021), both PAS and accuracy were modelled with psychometric curves with stimulus duration on the x-axis. Here, this was both not possible for all datasets (because they did not always include multiple stimulus durations), nor answering exactly the question of interest. It remains a possibility though that the relationships are not linear and so a linear model might not explain the data well – that case would be reflected in low indices of variance explained. To anticipate the section below, Q5 further aims to assess if individual participants' answers show linearity. Finally, PAS was not treated or modelled as ordinal, again with the purpose of facilitating comparisons with previous literature where the PAS was not assumed to be ordinal – e.g., direct mention of the assumption of PAS being an interval scale (Sandberg & Overgaard, 2015), 'mean PAS' being computed/reported/referred to (e.g., Overgaard & Sandberg, 2021; Sandberg et al., 2011; Schwiedrzik et al., 2009, 2011). Whether this assumption is challenged or not would require further research.

There was extremely strong evidence for both main effects (PAS: $BF_{alt} = 1.66 \times 10^{331}$, ± 0.01%; Study: $BF_{alt} = 1.18 \times 10^{11}$), and an interaction, $BF_{alt} = 4.40 \times 10^{46}$. The strongest model was the model with both predictors but no interaction, $BF_{alt} = 1.28 \times 10^{378}$ ± 1.57%), which was preferred over the full model (main predictors and interaction) by a factor of over 10^8. Adding Participant as a random effect left the pattern unchanged, except that the strongest model was the full model. Overall, PAS rating explained 54.7% of variance (SD = 1.33%, frequentist adjusted R^2 = 54.8%). The predictor Study explained around 5% of the variance: Bayesian R^2 = 5.34% (SD = 0.97%, adjusted R^2 = 4.14%). The BFs for the effect of PAS rating on accuracy (not rescaled) from regressions on each study are displayed in Figure 20. R^2 values are included in Table 11. In these models, the only known source of non-independence came from the two studies in Jimenez et al. (2019), which used the same participants, albeit in different tasks. Removing both studies left the pattern unchanged. Overall, these results highlight that a linear model explains moderately well the link between PAS and accuracy, thus suggesting that at the group level accuracy increases linearly with increasing PAS. However, the degree to which changes in the subjective measure follow the ones in the objective measure differs considerably across studies.

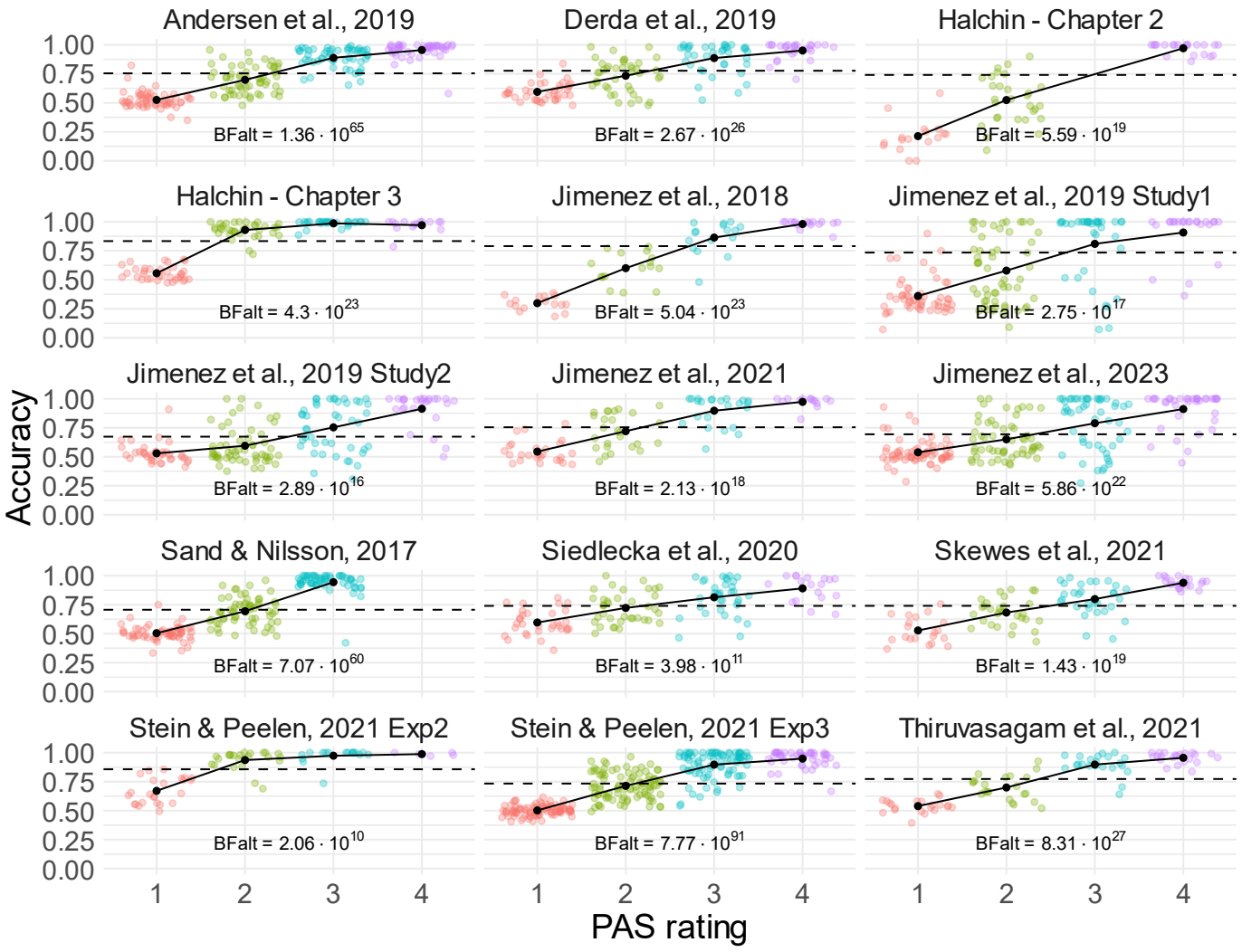


Figure 20. Distribution of mean accuracies at each PAS level for each participant. The black dots represent the grand means, and the dashed horizontal lines represent the overall accuracy across PAS ratings.

4.3.5 Q5. Is the change in PAS ratings gradual, or all-or-nothing?

The next question was whether changes in PAS are gradual or all-or-nothing, using mean accuracies for each participant and PAS level (i.e., data from Q4, not rescaled). Only participants who had data in all the PAS levels measured in each study were included, hence the lower number of datapoints in Figure 21 and the lower BFs in Table 12, compared to Figure 20. Data from Chapter 2 was not included in any analysis, given that no participants

had enough trials rated with PAS3 to allow inclusion, so only one comparison could be run (PAS1 vs PAS2).

Two types of comparisons were conducted. The first analysis compared group-wise accuracy scores at each PAS level in each study, with Bayesian paired t-tests (i.e., PAS1 vs PAS2, PAS2 vs PAS3, and PAS3 vs PAS4). Under the assumptions that accuracy is a good indicator of perceptual experience and that all participants used the scale in the same way, if awareness was all-or-nothing, then it would be expected that only one of these comparisons was supported by evidence. Conversely, if evidence supported the hypothesis of a difference for more than one comparison, then one interpretation would be that the shift was not all-or-nothing. However, another interpretation could be that participants do not use the scale in the same way, and that for each participant only one shift occurred, albeit at different locations along the PAS. Such a pattern, consistent with an all-or-nothing explanation, would be masked by a seeming graduality in group analyses. Hence, for the second analysis, linearity scores were computed for each participant, by identifying the (sign-independent) largest differences between two steps of the PAS (D_{max}), and dividing the sum of the other two scores (D_{sum}) by D_{max}. This method allows summarizing the pattern at the participant level in a single value, and comparing scores in each study to 0 in Bayesian one-sample t-tests. If the increase was gradual, then the differences between PAS1-PAS2, PAS2-PAS3, and PAS3-PAS4 (if available) would be similar, so the linearity score would be close to 2 (or 1, for three-level scale in Sand & Nilsson, 2017). Conversely, if there was only one substantial increase and minimal changes in the other two steps, the score would be close to 0. Linearity scores can also be negative if either the D_{max} or the D_{sum} are negative, suggesting that participants' performance decreased between PAS levels. Plausible explanations for these negative values could be noise in the data or participants misunderstanding the instructions (i.e., treating PAS as an index of confidence and using higher ratings to indicate that, for example, they had an 'almost clear' lack of experience of the stimuli). It could also be the case though that there was a mismatch between the stimuli and what the participants thought they saw.

The full results for both comparisons can be seen in Table 12. For the first comparisons (group-wise), in 13 out of 14 (92.9%) included studies there was at least moderate evidence for a difference between PAS1 and PAS2 compared to the null hypothesis, as well as

between PAS2 compared to PAS3. Finally, for PAS3 compared to PAS4, there was at least moderate evidence of a difference for 9 out of 13 (69.2%) studies that used a 4-level scale. These patterns might therefore seem to be consistent with the explanation that changes in subjective experience as indexed by the PAS occurred in a gradual manner.

To further assess if this conclusion is consistent at the participant level, linearity scores were compared. 4.5% participants had a negative D score. In 11 out of 14 studies (78.6%) there was at least moderate evidence that the linearity scores differed from 0 – in the remaining three, the evidence was inconclusive, likely because of the very low number of participants that passed the inclusion criteria (Figure 21). However, there were also substantial individual differences both within and across studies, with some participants' scores being around 0 (consistent with all-or-nothing), while others being towards ceiling (consistent with linear increases), even in the same sample. These findings highlight that group-based conclusions are not fully aligned with individual-level patterns.

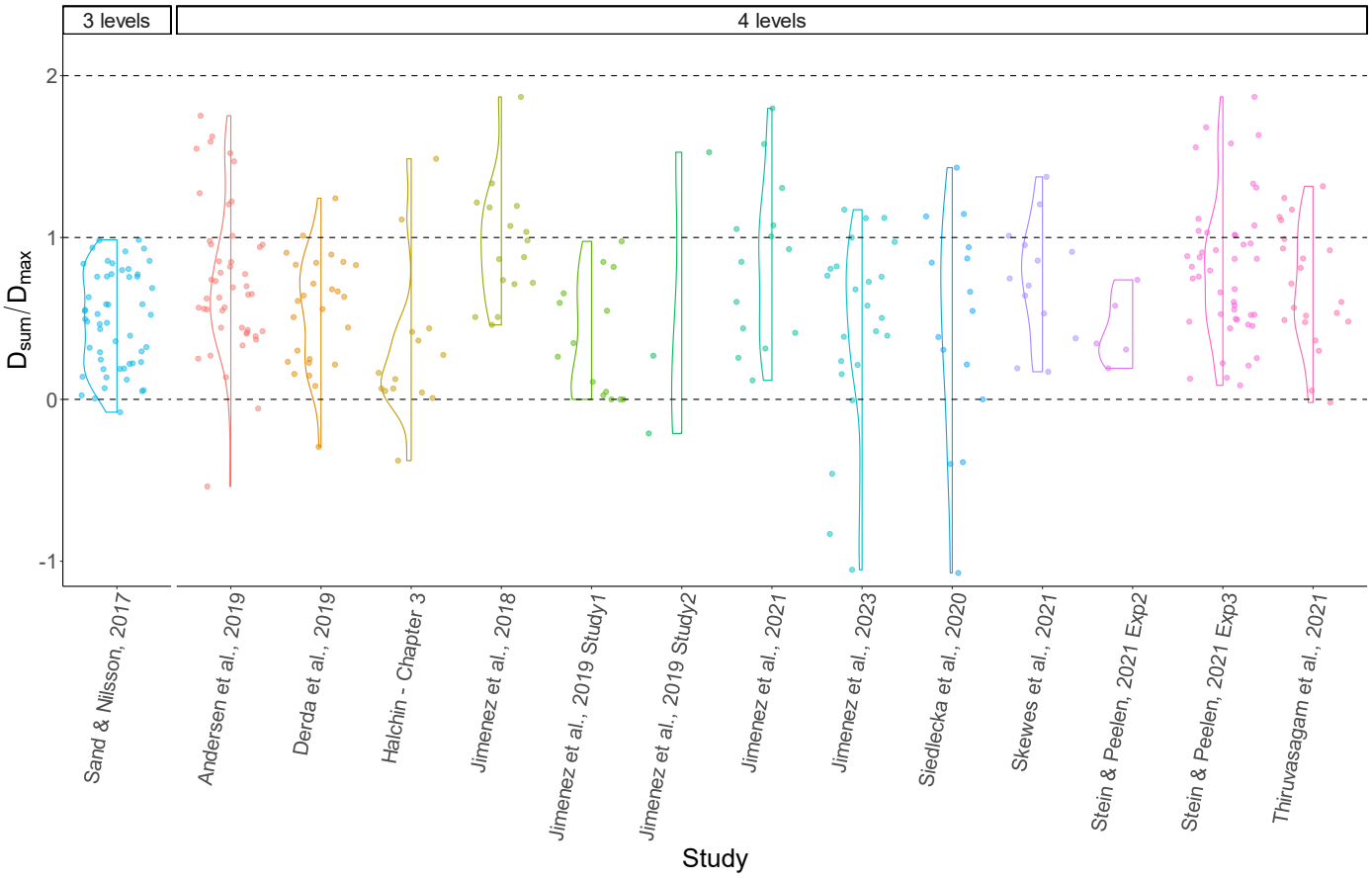


Figure 21. Linearity scores for each participant in each study. The dashed line marks 0 (an 'all-or-nothing' pattern), and limits for fully linear patterns (1 for 3-level PAS versions, 2 for 4-level versions).

Table 12. Individual comparisons between PAS ratings, in each study. The same notation convention as in Table 11 applies. All error terms under 0.2%.

	Study	BF PAS1 vs PAS2	BF PAS2 vs PAS3	BF PAS3 vs PAS4	BF linearity score vs 0
1	Andersen et al., 2019	$BF_{alt} = 8.52 * 10^{8}$ ++	$BF_{alt} = 1.38 * 10^{19}$ ++	$BF_{alt} = 2.18 * 10^{5}$ ++	$BF_{alt} = 1.79 * 10^{11}$ ++
2	Derda et al., 2019	$BF_{alt} = 8.77 * 10^{3}$ ++	$BF_{alt} = 6.95 * 10^{3}$ ++	$BF_{alt} = 65.8$ ++	$BF_{alt} = 4.94 * 10^{5}$ ++
3	Halchin – Chapter 3	$BF_{alt} = 2.80 * 10^{6}$ ++	$BF_{alt} = 1.06 * 10^{2}$ ++	$BF_{null} = 3.48$ *	$BF_{alt} = 2.17$
4	Jimenez et al., 2018	$BF_{alt} = 4.37 * 10^{5}$ ++	$BF_{alt} = 3.25 * 10^{6}$ ++	$BF_{alt} = 66.6$ ++	$BF_{alt} = 4.72 * 10^{5}$ ++
5	Jimenez et al., 2019 - Study 1	$BF_{alt} = 1.04 * 10^{3}$ ++	$BF_{alt} = 17.5$ ++	$BF_{alt} = 4.31$ +	$BF_{alt} = 23.3$ ++
6	Jimenez et al., 2019 - Study 2	$BF_{null} = 1.52$	$BF_{alt} = 1.69$	$BF_{alt} = 1.31$	$BF_{null} = 1.50$
7	Jimenez et al., 2021	$BF_{alt} = 2.56 * 10^{2}$ ++	$BF_{alt} = 3.66 * 10^{3}$ ++	$BF_{alt} = 15.9$ ++	$BF_{alt} = 7.5 * 10^{2}$ ++
8	Jimenez et al., 2023	$BF_{alt} = 1.17 * 10^{2}$ ++	$BF_{alt} = 66.2$++	$BF_{alt} = 9.45$ +	$BF_{alt} = 28.2$ ++
9	Sand & Nilsson, 2017	$BF_{alt} = 3.24 * 10^{12}$ ++	$BF_{alt} = 1.05 * 10^{19}$ ++	*NA*	$BF_{alt} = 2.31 * 10^{13}$ ++
10	Siedlecka et al., 2020	$BF_{alt} = 11.7$ ++	$BF_{alt} = 1.68 * 10^{2}$ ++	$BF_{null} = 1.66$	$BF_{alt} = 2.60$
11	Skewes et al., 2021	$BF_{alt} = 20.2$ ++	$BF_{alt} = 5.63 * 10^{2}$ ++	$BF_{alt} = 1.54 * 10^{3}$ ++	$BF_{alt} = 2.48 * 10^{3}$ ++
12	Stein & Peelen, 2021- Experiment 2	$BF_{alt} = 17.9$ ++	$BF_{alt} = 3.12$ +	$BF_{null} = 1.28$	$BF_{alt} = 6.19$ +
13	Stein & Peelen, 2021- Experiment 3	$BF_{alt} = 4.52 * 10^{9}$ ++	$BF_{alt} = 4.04 * 10^{15}$ ++	$BF_{alt} = 1.57 * 10^{6}$ ++	$BF_{alt} = 1.14 * 10^{13}$ ++
14	Thiruvasagam & Srinivasan, 2021	$BF_{alt} = 10.2 * 10^{4}$ ++	$BF_{alt} = 1.4 * 10^{6}$ ++	$BF_{alt} = 52.6$ ++	$BF_{alt} = 9.36 * 10^{5}$ ++

4.4 Discussion

Many research questions in the scientific study of consciousness require quantifying participants' conscious experience in specific instances, which is usually achieved through objective tasks such as making a judgement about a stimulus feature, or subjective tasks rooted in introspection. However, how these measures relate to each other and what they might inform about consciousness is not well-understood. This chapter investigated in more depth how a widely used subjective measure of awareness (the PAS) is linked to another metric argued to vary with awareness (performance in objective tasks), focussing on two dimensions: exhaustiveness (Q1-Q3) and sensitivity (Q4-Q5).

Regarding exhaustiveness, the same approach previously used in the literature was followed (e.g., Andersen et al., 2019; Sandberg et al., 2010). It was found (Q1) that in 7/15 studies there was at least moderate evidence favouring the hypothesis that participants' objective accuracy was above-chance when they reported no subjective experience (PAS1). Only three studies showed moderate evidence for the null, with the rest remaining either weak or inconclusive. Therefore, under the criterion previously set, these findings indicate that it would be misleading for a researcher to assume a priori that PAS will be exhaustive in their task. Further analyses into what factors might influence exhaustiveness (Q2) showed that stimulus duration influenced accuracy in PAS1 trials only in tasks requiring high-level judgments (such as whether a digit is odd or even), where the lower the stimulus duration the lower the accuracy, and evidence favoured it as predictor over the next strongest model by an order of magnitude. However, both stimulus duration and the full model only explained under 4% of the variance in accuracy, therefore suggesting that the overall contribution of the timing parameters is very small, and consequently that most of the variance is driven by currently unknown factors. A possible complication arises from the argument that maybe participants had more conservative response strategies in their PAS answers in studies where evidence for above-chance performance was found – so the 'No Experience' trial bins were contaminated by trials with weak awareness that still enabled above-chance performance. However, as discussed in Chapter 1, such possibility is present in all subjective reports, not only the PAS. Also, as Wierzchoń et al., (2014) discuss, even if participants were risk-averse and used the 'No Experience' answer too frequently as a response strategy, this could not be inferred from comparisons with an objective measure,

since it is not possible to estimate what the appropriate frequency of such answers would be if participants did not use any strategies.

Timing parameters played an even smaller role in the variance in accuracy at PAS1 in low-level tasks. Here, the evidence favoured the null hypothesis over all the models, moderately for stimulus duration and interaction, and weakly for mask duration. These findings are at odds with those of Sandberg et al., (2010), who reported that in their shape identification task, stimulus duration had a substantial and non-linear influence on accuracy at PAS1. Nevertheless, while no recommendations of design choices can be made for improving PAS exhaustiveness, one interesting pattern can be observed though in Figure 18: in low-level tasks, all tested stimulus durations yielded above-chance mean accuracies, while for high-level tasks durations up to around 40ms seemed to generate means close to chance. This observation likely explains why the only three studies that showed moderate evidence for the null in Q1 used high-level tasks (discrimination of face orientation, Stein & Peelen, 2021, Experiment 3; discrimination of words, Sand & Nilsson, 2017; identifying the content of natural scene images, Chapter 2 in this thesis) and primarily short stimuli durations. However, there was also high heterogeneity in accuracies at each stimulus duration, tentatively suggesting that other factors that were not tested in the current study could affect this pattern further. For instance, possible influences of specific stimulus categories (e.g., faces, arrows, letters, numbers etc.) and tasks (e.g., odd/even, left/right, upright/inverted etc.) were not tested here, because of too high heterogeneity in the available data. More available data, or experiments that systematically vary these factors, would be needed to further understand what might have driven the variance.

As for sensitivity (Q4-Q5), a clearer pattern emerged. There was extremely strong evidence that increases in PAS are accompanied by increases in accuracy (Q4), for all studies included. Interestingly though, the degree of variance in accuracy explained by PAS ranged substantially (from 39% to 84%) with studies in the lowest half of this range tending to be low-level tasks (4 out of 7 studies with lowest variance explained). More data would be needed though to assess whether task types influence the mapping between PAS and accuracy. In any case, this result is not unequivocal about whether such changes occur in a graded or all-or-nothing manner. When analysed at a group level, there was evidence that accuracy changed substantially both between PAS1 – PAS2, and PAS2 – PAS3, in 93% of the

included studies (Q5). However, at the individual level, linearity scores computed for each participant found substantial differences between participants' pattern of change, with most participants falling in-between the scores associated with all-or-nothing and gradedness (Figure 21). Each study also varied in terms of the range and spread of these scores. Whether these patterns can be attributed to true individual differences in awareness changes or other causes would require further research beyond group analyses. Nevertheless, these findings raise complications for attempting to answer whether awareness is graded or all-or-nothing. From a qualitative approach, the individual-level patterns were not consistent with either explanation, but this conclusion would be difficult to verify quantitatively, because there are no clear cutoffs for what would constitute either graded or all-or-nothing. Consequently, it is difficult to interpret these findings from the perspective of different theories of awareness, or how they might fit with previous findings from group analyses regarding gradedness.

As anticipated in 4.1, tackling different aspects of the relationship between PAS and objective measures can be seen as seeping into the discussion of whether PAS is a valid measure of awareness (Skóra et al., 2021; also referred to as the problem of 'coordination', Michel, 2019). The concepts used here and in the previous literature might seem to map somewhat onto formally identified aspects of validity (e.g., Adcock & Collier, 2001; Heale & Twycross, 2015). Sensitivity could be seen as criterion validity, or how well a specific measure relates to an existing benchmark measure of the same concept – more specifically convergent validity when expecting a high correlation. Exhaustiveness could be seen as content validity, or whether the measure comprehensively covers the target concept – in other words, if it can identify all cases of conscious perception. A different aspect of PAS validity, namely construct validity or whether the measurement outcome (i.e., PAS rating) allows drawing inferences about the target concept (i.e., clarity of experience), has also been discussed and criticized (Irvine, 2012, 2013). However, it is difficult to extend conclusions about exhaustiveness and sensitivity (in a statistical analysis sense, as described above in section 4.1 by Andersen and colleagues, 2019) to conclusions about validity. As Skóra and colleagues (2021) highlight, attempts to assess PAS validity are accompanied by assumptions rooted in theoretical positions about the target concept, and therefore conclusions about validity can only be drawn if one accepts these assumptions.

Whether studying the relationship between PAS and objective measures can inform about PAS validity therefore depends on whether one accepts objective measures as the benchmark for awareness – which comes with its own complications. Firstly, it is debatable whether objective measures provide a good index (i.e., whether they target aspects of awareness sufficiently similar with those of PAS) to assess validity against. Sandberg and Overgaard (2015) recognize that this is an assumption that they accept because the field does not yet have the methods or knowledge to contradict it. Secondly, it assumes that above-chance performance in trials labelled with PAS1 ('No Experience') can only be due to stimulus awareness – therefore not considering that it might be due to unconscious processing, in blindsight-like effects. Blindsight-like effects rely on the existence of a dissociation in which the measure of processing is above-chance whilst the measure of awareness indicates that participants did not consciously perceive the stimuli (although see Reingold & Merikle, 1988, for an alternative to this dissociation, requiring only relative differences between the two measures). These stances are mutually exclusive: one cannot interpret above-chance performance in PAS1 trials as both a benchmark for awareness and an index of unconscious processing. Taking the latter stance would allow interpreting the results in Q1 (above-chance performance in PAS1 trials in almost half of the included studies) as showing, in fact, strong evidence for unconscious perception. However, it would also entail accepting that none of the previous conclusions about PAS content validity drawn from failures to find above-chance performance are valid, because such results can only be interpreted as failures to find unconscious processing. Finally, as discussed in Chapter 1, equating objective performance with awareness also implies that the only explanation for chance performance is lack of awareness, and by extension that it is not possible to have awareness of a stimulus and still perform at chance. In other words, it assumes that it is not possible to have phenomenal consciousness without access consciousness (P-without-A), an assumption contested by both theoretical (e.g., Block, 1995) and empirical (Amir et al., 2023) work.

Ultimately, it might not even be possible to explain if above-chance performance is due to low exhaustiveness or unconscious processing, a caveat that Sandberg and colleagues (2010) also recognize. To circumvent the problem of relying on objective measures to test content validity/exhaustiveness, one could manipulate factors argued to influence the quality of

experience, such as stimulus duration or SOA. This approach would require making strong assumptions a priori that selected parameters would result in an 'unconscious' condition, which may or may not be justified given the specific experimental design. Then, one could compare the distribution of PAS answers in such condition with that of catch trials where nothing is presented, and infer no subjective awareness should no difference be observed. I emphasize here, as shown in Q3 (section 4.3.3) and Figure 19, that it should not be assumed that 100% of stimulus-absent catch trials would be rated with PAS1, and that the frequency is consistently lower, on average around 85%. The factors that might drive PAS2+ answers in stimulus-absent trials are diverse, and could be experience-independent participant mistakes (i.e., accidental button presses, incorrect understanding of instructions). Speculatively, they might also be experience-dependent: it could be argued that participants might have experienced internally generated illusions in line with their prior expectations about the stimuli (e.g., Chalk et al., 2010), thus leading to PAS2+ answers. However, these factors do not invalidate the suggested comparison, because they do not capture awareness of a presented stimulus. Nevertheless, such comparison might not be suitable for all research questions and would need to first be examined empirically.

One important consideration that this study does not address though is whether the variability in phrasings and/or number of levels in the PAS affects the ratings independently of the task. This is because almost all the datasets in this analysis used a 4-point scale, that starts at 1 (as opposed to 0, e.g., Peremen & Lamy, 2014), and refer to the whole stimulus rather than the task-relevant features. However, these choices are not always the case in the literature; regarding the 4-scale PAS, some studies adopt a 3-point (e.g., Eklund & Wiens, 2018; Sand & Nilsson, 2017) or even continuous (e.g., Stein & Peelen, 2021, Experiment 2; Wierzchoń et al., 2019) version of the scale. Regarding the latter choice, making the whole stimulus the target of the clarity judgment is a departure from the original study (Ramsøy & Overgaard, 2004), which specifically targeted features (shape, colour, position). This choice is implicitly or explicitly linked to a theoretical stance on what should constitute a criterion for 'lack of awareness': no awareness of the task-relevant features only (as argued by Breitmeyer, 2015; Dienes & Seth, 2010b), or no awareness of the whole stimulus. Again, whether such choices result in substantial differences in ratings is an empirical question that, to my knowledge, has not been addressed so far. Therefore, this would be a worthwhile next

step, since loosely justified changes in an existing scale can be deemed ‘questionable measurement practices’ (Flake & Fried, 2020).

Altogether, PAS has proven useful in prompting literature-wide considerations about how researchers evaluate awareness, and how the concept of ‘perceptual clarity’ may differ from confidence or other meta-cognitive judgments. However, it remains unclear how looking at the relationship with objective measures might inform about the suitability of the PAS as an awareness measure, or whether this question can be answered with the current available methods. Some steps forward would be to continue to design experiments where PAS is collected alongside other measurements, and to continue to make experimental data and code publicly available, with the long-term goal of facilitating understanding of how perceptual clarity judgments relate to other types of judgments and to the theories and conceptualizations of awareness.

5 Chapter 5

5.1 General summary

This thesis had three aims: firstly, to reassess, using improved experimental designs and Bayesian statistical analyses, previous findings of VPL from visual information that participants had no awareness of. Secondly, to evaluate how different experimental factors, in particular those related to measurement, influence conclusions about learning. Finally, to assess the relationship between objective task performance and answers on the Perceptual Awareness Scale, a popular measurement of clarity of subjective experience.

To summarize the main results, in Chapter 2, the findings of improvements from Pre to Post-Exposure (2.5.2) in the Unconscious and Mostly Unconscious conditions were observed only in Experiment 2, where participants were intractably exposed to the correct answer of what the image depicted, in the MCQ paradigm. No such effect was consistently observed in Experiment 1, where identification was collected through free-naming, so the only source of information about the two-tone content was from the images. In Chapter 3, the findings of improvements from Day 1 to Day 3 in all three measurements tested (discrimination, detection, and PAS) extended both to the condition that trained on sub-threshold contrasts and to the condition that did not complete the training, highlighting that the training could not have been the source of the improvements. In Chapter 4, the evidence highlighted that the criterion of exhaustiveness (chance performance in trials rated with PAS = 1, or “No Experience”) was not met in a large percentage of the studies in the sample. On the other hand, the criterion of sensitivity was met, as all studies showed strong evidence that PAS ratings increase with task performance.

5.2 VPL without stimulus awareness?

The findings from both Chapters 2 and 3 converge to the broader conclusion that there is no evidence for a learning effect following exposure to unconscious stimuli that can only be explained by the exposure. This broader conclusion is complemented by the previous results claiming VPL from unconscious stimuli, presented in Chapter 1 (1.3), arguably not being equivocal either. Seitz et al. (2009, Experiment 2) concluded that participants were not aware of the orientation stimuli during training based on a separate session, conducted after

the end of the experimental sessions. In this awareness test session, participants were asked to make an orientation response only if they detected the gratings presented under CFS; unawareness was concluded based on the findings that in most trials participants made no answers (implying they did not detect the stimuli) and when they did answer, their accuracy was not different from chance. However, the design and analysis of the awareness test do not warrant the conclusion that participants had no awareness during training. First, lack of awareness was inferred from a frequentist non-significant result and no tests for assessing evidence for the null were reported, so it is possible that the evidence was insensitive to whether there was a difference from chance – especially since participants did not provide answers in the majority of trials (around 80%). Secondly, the awareness test introduced a task compared to training (where there was no task) and removed the rewards, thus creating fundamentally different conditions – it could be that participants' low objective performance in the awareness test stemmed from them not being motivated or incentivised to do the task, thus leading to the incorrect conclusion that they lacked awareness during training. A second awareness test, which found the same non-significant result, was conducted with rewards but in a different group of participants who did not undergo training. This second test is again an unsuitable benchmark: in addition to the same analysis fallacy as the first test, inferring awareness in one group of participants based on a different group of participants is problematic, because it assumes that the effectiveness of CFS is the same across groups. However, CFS effectiveness was found to be affected by individual differences (Blake et al., 2019), and the small sample sizes (n = 8 for the awareness test, n = 4 for Experiment 2) in Seitz et al. (2009) makes the study's conclusions more susceptible to possible heterogeneity from individual differences.

Nishina et al. (2007) also concluded that participants had no awareness of the gratings-in-noise during the training. However, in both Experiments 1 and 2, the contrast of the gratings was the same for all participants, picked from a psychometric function fit on means across participants in a pilot experiment. This approach means that some participants in the small sample (n = 7 in Experiment 1, n = 9 in Experiment 2) had above-chance accuracy – and therefore, in line with their criterion of unawareness based on discrimination accuracy, they were aware of the gratings. This detail was directly mentioned by the authors, that in both experiments most but not all participants had chance performance at the trained contrast.

Even if the learning effect is not explained by awareness during training, it is not necessarily explained by the training either. In Experiment 1, the training contrast was 12% and yielded on average chance performance. The lowest contrast included in the pre-post training measurements was 15%, which had slightly above-chance performance pre-training. However, in the only condition where a learning effect was found, there was no improvement at the 15% level (based on supplementary figure 2, contrast-specific comparisons were not reported), but there was an improvement at all the other contrast levels that already yielded performances much higher than chance before training. In Experiment 2, the training contrast was also the lowest contrast tested pre- and post-training, at 15%. A similar pattern of findings was reported there (supplementary figure 4): of the three conditions where a learning effect was found, in only one performance at 15% was higher in post-training; in the other two, performance was either the same, or lower in post-training. Altogether, this pattern does not seem consistent with the explanation that the learning stemmed from the unconscious stimuli during training, because robust increases at the trained parameter would have been expected. The findings from Pascucci and colleagues (2015), from a very similar TIPL paradigm to Nishina et al., (2007) which trained participants with gratings of subthreshold contrast (12%), showed the same pattern in the graphs: a pre-post training general reduction in contrast PF threshold of identifying which of two intervals contained a grating, but small or no changes at the lowest value tested (10% contrast). However, the authors did not report specific comparisons for accuracy at the trained contrast with specific evidence for the null, so it is not possible to empirically assess if this interpretation is correct. Nevertheless, applying the definitions of VPL and awareness discussed in Chapter 1 and being mindful of statistical concerns, there seems to be no reliable evidence in the current literature of a VPL effect led unambiguously by stimuli that participants claimed they lack awareness of, regardless of the complexity of the task.

Following up on the discussion in Chapter 3 (section 3.5), one possible criticism of this conclusion and any null finding against unconscious VPL is simply that the training was not sufficiently long. If the masked templates/greyscale images in Chapter 2 were presented – albeit masked – more than once, or if the training in Chapter 3 would have involved more than 1000 trials, then one can speculate that, in theory, learning might have occurred. While I agree with this theoretical possibility, the issue is intractable. What is the upper limit of the

number of unconscious exposures by which it can be concluded that learning requires consciousness? For both study designs, one could draw comparisons with training protocols where the training is conducted on perceived stimuli, but this approach would not be conclusive either – if unconscious learning does not occur over a number of trials that generate learning when consciously experienced, at most it would allow the conclusion that unconscious learning does not occur on the same timeline as learning from consciously experienced stimuli. This conclusion cannot be drawn from the current work, but it could be a promising avenue for future research.

5.3 Measuring 'unawareness' of visual information

One focal point addressed throughout the thesis is whether the choice of index used to delimit trials/conditions where participants had no awareness influences conclusions about the absence of awareness, and possible effects of the unconscious information.

The evidence discussed in the thesis suggests that while generally consistent, different indices of unawareness do not always lead to the same conclusions about the presence of a learning effect. In Chapter 2, for the measure of subjective meaningfulness, there was no advantage of exposure to congruent templates regardless of how unawareness was defined during the exposure stage (based on the SOA alone, PAS alone, identification accuracy alone, or combinations of these factors with different degrees of conservativeness), in neither Experiment 1 nor 2 (sections 2.5.1 and 2.6.1). Although most comparisons also indicated that there was no advantage in identification accuracy either, solely indexing based on incorrect answers or PAS = 1 ("No Experience" of the content) in fact yielded inconclusive results, but only in one of the two methods of probing identification accuracy. Table 5 (section 2.7.3) provides further examples of discrepancies in conclusions. Inconclusive results entail that the data presents patterns that are not inconsistent with the explanation that a difference could be present – which is fundamentally a different conclusion from support for the null hypothesis. Two broader points become apparent. The first is the importance of using statistical analyses that allow quantifying the evidence for the null, such as Bayes Factors – under typical frequentist analyses, it would have been impossible both to disentangle these patterns, and more generally to conclude the absence of an advantage.

The second is that the landscape seems more complicated: whether different definitions of 'unconscious' lead to the same conclusions about the target effect (i.e., 'unconscious' learning) might thus depends on experimental design factors such as the operationalization of the target effect (i.e., subjective or objective indices of learning) and how it is measured (i.e., free-naming or MCQs). To some extent, this conclusion echoes findings from Chapter 4 as well, if one chooses to interpret above-chance task performance in PAS1 trials as evidence of unconscious processing. There (Q1, section 4.3.1), the same index of unawareness (namely PAS1) was associated with vastly different task performances across different studies (and consequently, different stimuli, tasks, etc.), with evidence spanning all possible conclusions (inconclusive, chance performance, above-chance performance). Moreover, experimental design factors like stimulus duration and task type influenced the accuracy of these trials (Q2, section 4.3.2). Altogether, the emergent pattern is that generalizability of findings can be very poor across experimental designs, regardless of whether the same indices of awareness are used. In turn, the use of experimental findings to guide the development of theories of consciousness or to map the boundaries of 'unconscious' processing becomes severely limited.

In any case, if different definitions of unawareness can lead to different conclusions about target effects, then one possible explanation could be that each definition might capture different 'types' (or facets) of unawareness, which might not be compatible. In Chapter 2, only qualitative comparisons between definitions were possible, because all indices were on different scales: there is no way to determine if participants' conscious experience was the same when they answered "No Experience" of the content on the PAS as when they responded incorrectly or were exposed to the images in the Short SOA condition. Looking at the overlap in the distribution of trials could be one aspect to consider, but as shown by pilot data (Figure 7, consistent with the main experiments as well, figure not included), the overlap in trials is very variable across indices. However, this question was addressed in Chapter 3. There (Q4P, section 3.3.5), no difference was found between discrimination and detection performances at the trained contrast (i.e., yielding under 60% discrimination accuracy), suggesting by extension that discrimination unawareness (inability to reliably discriminate the direction of arrows) is the same as detection unawareness (inability to reliably respond if arrows were displayed at all). This pattern would correspond to a failure

of finding classical 'blindsight' in healthy individuals (other null findings: Balsdon & Azzopardi, 2015; Peters & Lau, 2015; Rajananda et al., 2020), usually operationalized as above-chance performance on discrimination of some stimulus feature without detection of stimulus presence. Interestingly though, the evidence also suggested that detection inflection points (i.e., contrast level yielding ~81% accuracy) were indeed higher than discrimination inflection points, for most participants (Chapter 3, Figure 13) – thus suggesting that at certain contrasts, participants' ability to respond correctly if the arrows pointed to the left or the right was better than their ability to respond correctly if the arrows were presented. More research would be needed though to understand what gave rise to the discrepancy in the current experiment, and at what contrast level the two indices might start to diverge in this task. Moreover, whether this pattern is indicative of blindsight (or 'Perception > Awareness' in Peters & Lau, 2015) is debatable. On the one hand, this finding might seem at odds with results from Peters and Lau (2015, control experiment), who found an almost-exact mapping between discrimination accuracy (identifying left/right orientation of gratings in a 2IFC task) and detection accuracy (percentage of trials in which participants said that the target-present interval contained the more visible stimulus). However, such conclusion would not be justified, given the substantial differences in task and the small, heterogeneous sample size (n = 3, with one participant displaying a pattern consistent with that in Chapter 3). On the other hand, Peters and Lau (2015) also found the same dissociation (higher discrimination performance relative to subjective detection, but not around the chance level), only when detection was inferred from betting (taken to indicate confidence, Experiments 1 and 2). While they did not attribute this result to 'Perception > Awareness' but rather differential noise between the discrimination and detection judgments in the 2IFC task, the mixed results across experimental designs further speak to the challenge of generalizing conclusions about perception without awareness.

5.4 Conclusions and future directions

The work conducted in this thesis demonstrated the importance of intentionally examining questions about unconscious learning from multiple methodological angles. It also demonstrated the importance of designing appropriate control conditions when studying unconscious learning, and highlighted new difficulties in the use of a popular measure of

subjective awareness. As discussed throughout this thesis, the plurality of methods in consciousness research raises substantial difficulties in comparing results across experiments and identifying patterns among results. Natural next steps for future research would therefore be to exercise caution in interpreting relevant findings about the target effect from paradigms with different experimental designs, and to continue to systematically investigate in which conditions, if any, unconscious learning exists.

One final consideration must be given to open scientific practices (e.g., preregistration, sharing data and research materials) and consciousness research. While engaging with open practices is not in itself an indicator of research quality (Devezer & Penders, 2023), it has become apparent that it has substantial scientific benefits, such as allowing higher scrutiny of research findings or allowing researchers to assess other aspects of the data that might be of interest to them but were not addressed in the published papers (Nosek et al., 2012). The process of obtaining Stage 1 In-Principle Acceptance for the research in Chapter 2 required me to engage with such practices myself, which ultimately improved the quality of the research. Moreover, the analyses and conclusions in Chapter 4, raising concerns and challenges about a popular subjective measure of awareness, would not have been possible without consciousness researchers valuing open practices and engaging with the substantial additional effort required by preparing and sharing materials (e.g., ensuring participant anonymity, compliance with data sharing policies, good structure and legibility of the shared materials) – effort currently not well-incentivised. Neither would have been the rectification of inconsistencies in published work, detected during the re-analysis for Chapter 4 (Andersen et al., 2023; Thiruvasagam & Srinivasan, 2023). More widespread adaptation of data and code sharing would therefore continue to enable future attempts to systematically compare results, and to assess the potential impact of different methodological choices on conclusions about consciousness. Because of these reasons, I argue that these practices would not only continue to be beneficial for consciousness research, but they should become standard practice. The hope is that, by adopting these approaches, there will be a broader wealth of publicly available data for systematic comparisons, which would, in turn, facilitate robust new insights in the study of the unconscious.

Appendices

Appendix 1 – PAS training (Chapter 2)

Modified description of the PAS used in all 4 experiments in Chapter 2 (Pilots 2 and 3, Experiments 1 and 2). The changes in the description of the scale were prompted by previous concerns that the scale might measure confidence instead of clarity of experience (Irvine, 2012).

1 (no experience) = I did not see the content of the image at all.

2 (a brief glimpse) = I had a feeling of seeing the content, but I did not know what it was.

3 (an almost clear experience) = I had seen the content of the image almost clearly.

4 (a clear experience) = I had clearly seen the content of the image.

Items on the PAS quiz:

Level Description	**Correct PAS Answer**
I had no impression of the content of the image.	1
I could not see the content presented.	1
I only had a feeling of seeing the content of the image.	2
I had a weak glimpse of the content of the image.	2
I saw the content mostly but not completely clearly.	3
I saw the content of the image with weak clarity.	3
I saw the content non-ambiguously.	4
I very clearly saw the content of the image.	4

Appendix 2 – Image details (Chapter 2)

Outline of the experiment design

Block no.	Image type	Block contents
Block 1	Two-tone	Two-tone 1-8, randomized
Block 2	Greyscale	Greyscale 1-8 + attention checks, randomized
Block 3	Two-tone	Two-tone 1-8 + two-tone 9-16, randomized
Block 4	Greyscale	Greyscale 9-16 + attention checks, randomized
Block 5	Two-tone	Two-tone 9-16 + Two-tone 17-24, randomized
Block 6	Greyscale	Greyscale 17-24 + attention checks, randomized
Block 7	Two-tone	Two-tone 17-24, randomized

The outline above includes the 24th pair of two-tone and greyscale images, which was added only to achieve equal block lengths, as it did not come from the validated stimulus set. This image was removed from the analyses, hence 23 pairs remained.

Low-level image properties and PAS

Additional quality checks of participants' use of the PAS tested whether participants' PAS answers were influenced by RMS contrast, luminance, edge density, and spatial frequency of the templates in Pilots 2 and 3. The results from separate Bayesian ANOVAs with default priors, PAS as categorical predictor, and low-level property as dependent variable, are included in the table below. Overall, these results suggest no relationship between PAS ratings and low-level image properties.

Property	**Pilot 2 Free Naming BF (+ error)**	**Pilot 3 (MCQ) BF (+ error)**
RMS contrast	BF_{null} = 9.46 ± 0.59%	BF_{null} = 1.42 ± 0.51%
Luminance	BF_{null} = 11.1 ± 0.6%	BF_{null} = 1.63 ± 0.52%
Edge density	BF_{null} = 26.8 ± 0.62%	BF_{null} = 1.89 ± 0.54%
1/f slope	BF_{null} = 8.57 ± 0.59%	BF_{null} = 24.7 ± 0.64%

Appendix 3 – Summary of Chang et al. (2016) experimental design (Chapter 2)

As illustrated in Chang et al. (2016), page 2, Figure 1, each two-tone was followed by a prompt to respond Yes or No to the question 'Can you recognize and name the object in the image?' (p.3). Each set of four two-tones was repeated twice per block and each block was repeated twice; the second block was followed by a free-naming identification task for the two-tone images only, in which all two-tones were presented again, followed by a free-naming identification task. The stimulus onset asynchrony (SOA) was calculated as 67ms, from the beginning of the greyscale stimulus until the beginning of the mask. This was followed by a noise mask for 1933ms and a prompt to respond to the same question presented in the two-tone trials. Each greyscale trial was presented twice before participants were asked to verbally name the two-tones Post-Exposure.

Appendix 4 – PAS training (Chapter 3)

Table with PAS quiz items. The items were randomized for each participant.

Scale Description	Correct PAS answer
I had no impression of the arrows.	1
I could not see the arrows presented.	1
I have only a feeling that the arrows were presented.	2
I had a weak glimpse of the arrows.	2
I saw some arrows better than others.	3
I saw the arrows with weak clarity.	3
I saw the arrows non-ambiguously.	4
I very clearly saw the arrows.	4

Appendix 5 – Psychophysics methodology details (Chapter 3)

Method of limits – procedure details

In half of the blocks the arrows' contrast progressively increased, and in the other half progressively decreased. The increasing blocks stopped when participants responded that they started to perceive the arrows, which was defined as both PAS above 1 and correct direction discrimination. Conversely, the decreasing blocks stopped when participants reported not seeing the arrows anymore, which was defined as PAS of 1 and an incorrect response. If a participant answered with PAS 1 but had correct discrimination, or other unexpected combinations (e.g., incorrect discrimination and PAS above 1), the block paused and they were asked to briefly justify their responses verbally, and the experimenter decided if the block would finish or continue. This was done to ensure their answers reflected their perception, rather than lapses or lucky guesses. The trials progressed in steps of 4 intensities. The first block of each kind started from the lowest visibility that could be presented, respectively close to maximum visibility. For the subsequent blocks of each kind, the starting contrast was 8 levels before the stopping contrast in the first block of each type, to minimize likely redundant trials. At the end of the blocks, the average of all stopping contrasts was taken, and a range between ± 6 from the mean was generated, to be used in the MoC.

Lookup table between the contrast level indices used and their corresponding values in cd/m^2

Contrast index	**cd/m^2**
0 (background)	34.08
1	33.44
2	32.85
3	32.18
4	31.58
5	30.91
6	30.37
7	29.68
8	29.16
9	28.48
10	27.95
11	27.38
12	26.78

13	26.19
14	25.72
15	25.17
16	24.72
17	24.16
18	23.62
19	23.17
20	22.64
21	22.19
22	21.73
23	21.23
24	20.84
25	20.32
26	19.84
27	19.45
28	18.97
29	18.52
30	18.11
31	17.71
32	17.29
33	16.88
34	16.5
35	16.08
36	15.66
37	15.28
38	14.92
39	14.59
40	14.25
41	13.96
42	13.61
43	13.23
44	12.88
45	12.53
46	12.23
47	11.9
48	11.58
128 (black)	0.15

Appendix 6 – Chapter 4

Links to open data and other supporting materials for the included studies

Andersen et al., 2019 - https://osf.io/ecxsj/

Derda et al., 2019 - https://osf.io/63tbu/

Jimenez et al., 2023 - https://osf.io/stcdh/?view_only=af3fe6f2ca8941a1b894b7a83fbd2234

Sand & Nilsson, 2017 - https://osf.io/fsu43/

Siedlecka et al., 2020 - https://osf.io/b4qk7/?view_only=f66822a713924b9e8fbb7973a69f43ac

Skewes et al., 2021 - https://osf.io/cfk43/

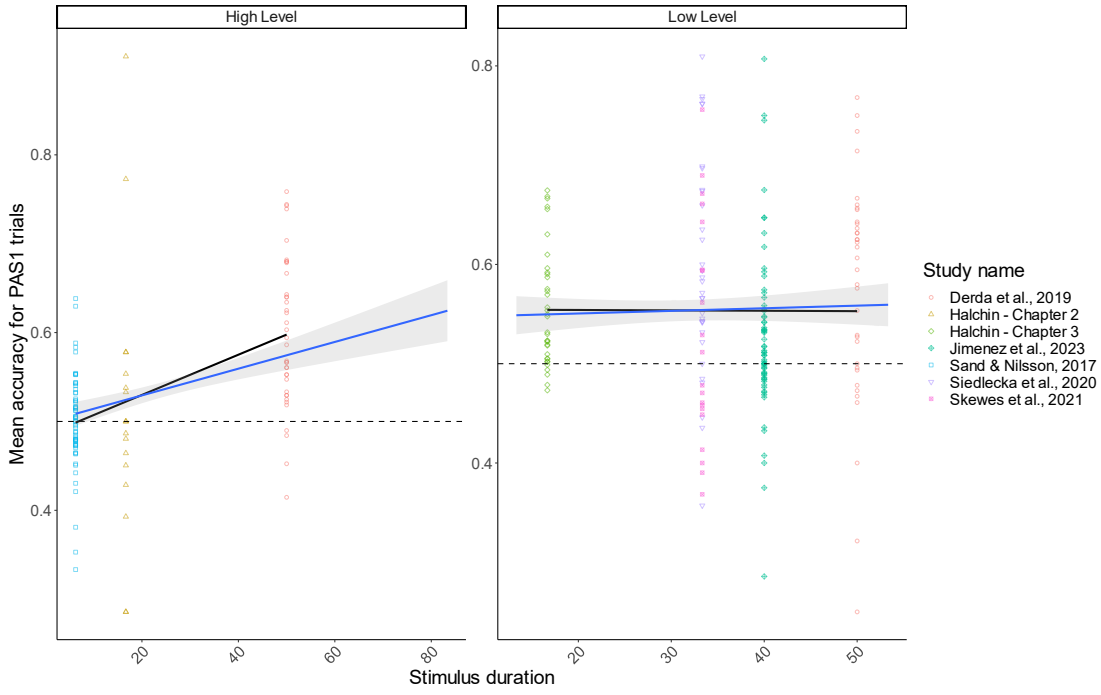
Stein & Peelen, 2021 - https://osf.io/sn8cr/ NB: trial-by-trial PAS data was kindly provided privately by T. Stein.

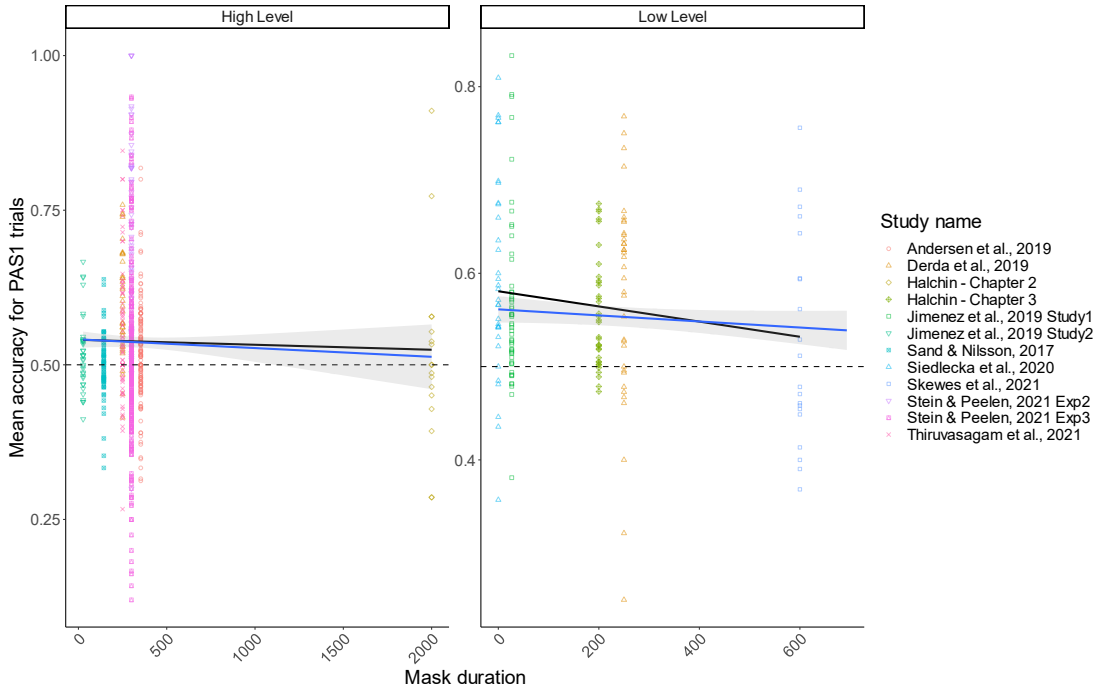
Thiruvasagam & Srinivasan, 2021 - https://osf.io/k5nf9/

Data for Jimenez et al., (2018, 2019, 2021) was kindly provided privately by M. Jimenez.

Figure S1. Q2 follow-up (Chapter 4)

Studies from Figure 18 that contain only one stimulus (or mask) duration only, with the black lines representing the regression lines. For comparison, the dark blue lines with grey ribbon are the regression lines (and SE) from Figure 18, on the full sample. BFs are included in the table below. As observed, removing the studies with non-independent data resulted in models that overlap with the SEs and overall comparable BFs, suggesting therefore that including datasets with non-independencies did not affect the overall pattern in Q2.





	Low Level tasks	High Level tasks
Stimulus duration	BF_{null} = 6.27	BF_{alt} = 1.63 * 10^5 ± 0.01%
Mask duration	BF_{null} = 10.9	BF_{alt} = 1.17

Figure S2. Q3 additional visualization

It is worth noting that Jimenez et al. (2023) excluded and replaced 24 participants for answering with PAS4 in more than 10% of blank trials.

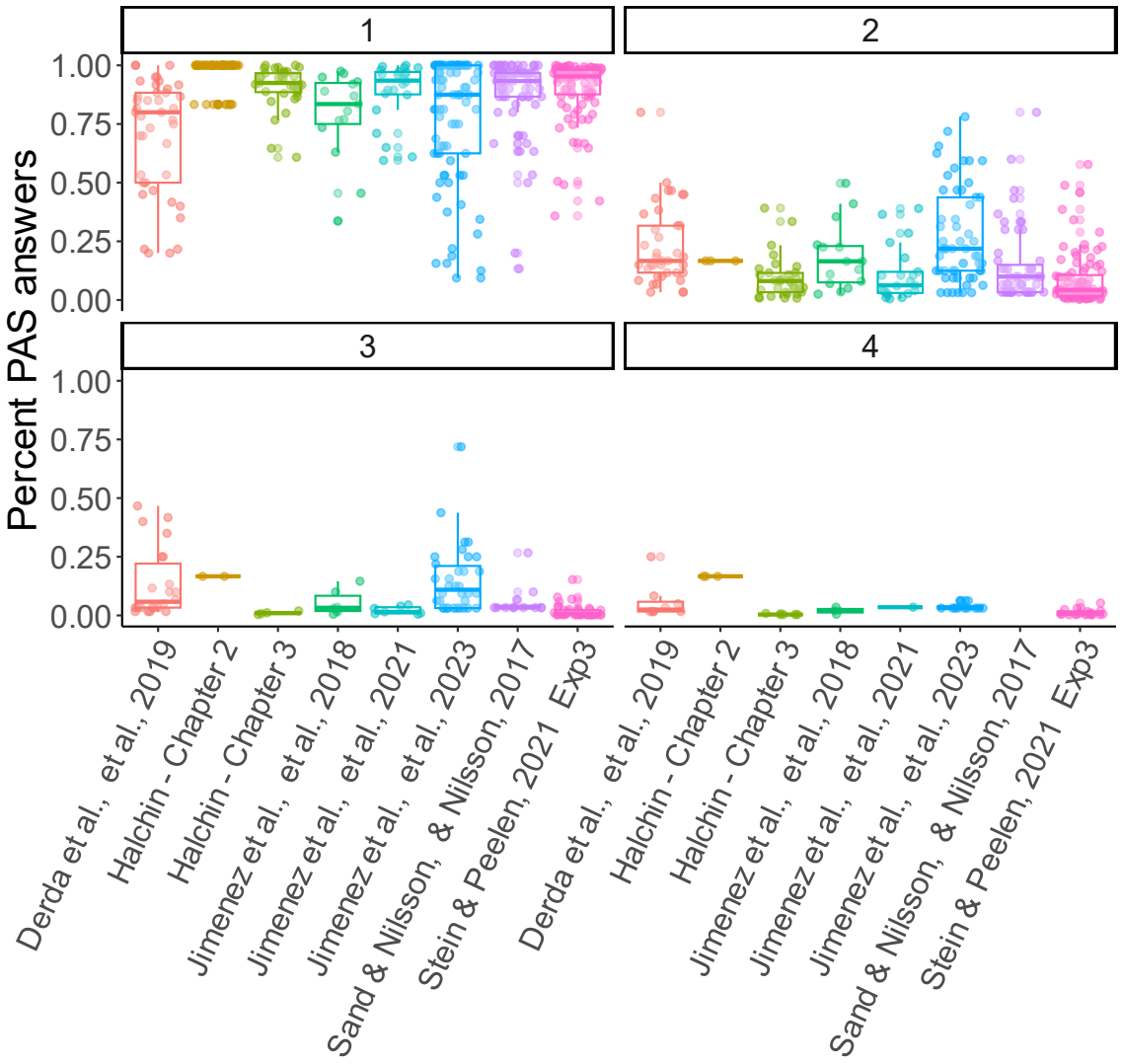


Figure S2. Distribution of the percentages of stimulus-absent catch trials answered with each of the PAS levels, for each participant. Panels represent PAS levels. No data is shown in Panel 4 for Sand & Nilsson (2017) because their PAS had three levels only.

References

Adcock, R., & Collier, D. (2001). Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review*, *95*(3), 529–546. https://doi.org/10.1017/S0003055401003100

Adolph, K. E., & Kretch, K. S. (2015). Gibson's Theory of Perceptual Learning. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* (pp. 127–134). Elsevier Inc. https://doi.org/10.1016/B978-0-08-097086-8.23096-1

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, *387*(6631), 401–406. https://doi.org/10.1038/387401a0

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, *8*(10), 457–464. https://doi.org/10.1016/j.tics.2004.08.011

Albrecht, T., Klapötke, S., & Mattler, U. (2010). Individual differences in metacontrast masking are enhanced by perceptual learning. *Consciousness and Cognition*, *19*(2), 656–666. https://doi.org/10.1016/j.concog.2009.12.002

Amerio, P., Michel, M., Goerttler, S., Peters, M. A. K., & Cleeremans, A. (2023). *Featural blindsight in a 2-Interval Forced-Choice task*. OSF. https://doi.org/10.31234/osf.io/jcae4

Amir, Y. Z., Assaf, Y., Yovel, Y., & Mudrik, L. (2023). Experiencing without knowing? Empirical evidence for phenomenal consciousness without access. *Cognition*, *238*, 105529. https://doi.org/10.1016/j.cognition.2023.105529

Andersen, L. M., Overgaard, M., & Tong, F. (2019). Visual expectations change subjective experience without changing performance. *Consciousness and Cognition*, *71*, 59–69. https://doi.org/10.1016/j.concog.2019.03.007

Andersen, L. M., Overgaard, M., & Tong, F. (2023). Corrigendum to “Visual expectations change subjective experience without changing performance” [Conscious. Cogn. 71 (2019) 59–69]. *Consciousness and Cognition*, *109*, 103479. https://doi.org/10.1016/j.concog.2023.103479

Axelrod, V., & Rees, G. (2014). Conscious awareness is required for holistic face processing. *Consciousness and Cognition*, *27*, 233–245. https://doi.org/10.1016/j.concog.2014.05.004

Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. In S. Laureys (Ed.), *Progress in Brain Research* (Vol. 150, pp. 45–53). Elsevier. https://doi.org/10.1016/S0079-6123(05)50004-9

Baars, B. J., & Franklin, S. (2007). An architectural model of conscious and unconscious brain functions: Global Workspace Theory and IDA. *Neural Networks*, *20*(9), 955–961. https://doi.org/10.1016/j.neunet.2007.09.013

Bacon-Macé, N., Macé, M. J.-M., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research*, *45*(11), 1459–1469. https://doi.org/10.1016/j.visres.2005.01.004

Balsdon, T., & Azzopardi, P. (2015). Absolute and relative blindsight. *Consciousness and Cognition*, *32*, 79–91. https://doi.org/10.1016/j.concog.2014.09.010

Balsdon, T., & Clifford, C. W. G. (2018). Visual processing: Conscious until proven otherwise. *Royal Society Open Science*, *5*(1), 171783. https://doi.org/10.1098/rsos.171783

Bao, M., Yang, L., Rios, C., He, B., & Engel, S. A. (2010). Perceptual Learning Increases the Strength of the Earliest Signals in Visual Cortex. *Journal of Neuroscience*, *30*(45), 15080–15084. https://doi.org/10.1523/JNEUROSCI.5703-09.2010

Bengtsson, H. (2022a). *R.utils: Various Programming Utilities* (2.12.0) [Computer software]. https://CRAN.R-project.org/package=R.utils

Bengtsson, H. (2022b). *R.matlab: Read and Write MAT Files and Call MATLAB from Within R* (3.7.0) [Computer software]. https://CRAN.R-project.org/package=R.matlab

Bhatt, R. S., & Quinn, P. C. (2011). How Does Learning Impact Development in Infancy? The Case of Perceptual Organization. *Infancy*, *16*(1), 2–38. https://doi.org/10.1111/j.1532-7078.2010.00048.x

Blake, R., Goodman, R., Tomarken, A., & Kim, H.-W. (2019). Individual differences in continuous flash suppression: Potency and linkages to binocular rivalry dynamics. *Vision Research*, *160*, 10–23. https://doi.org/10.1016/j.visres.2019.04.003

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Brams, S., Ziv, G., Levin, O., Spitz, J., Wagemans, J., Williams, A. M., & Helsen, W. F. (2019). The relationship between gaze behavior, expertise, and performance: A systematic review. *Psychological Bulletin*, *145*(10), 980–1027. https://doi.org/10.1037/bul0000207

Breitmeyer, B. G. (2015). Psychophysical “blinding” methods reveal a functional hierarchy of unconscious visual processing. *Consciousness and Cognition*, *35*, 234–250. https://doi.org/10.1016/j.concog.2015.01.012

British National Corpus Consortium. (2007). *British National Corpus, version 3 (BNC XML Edition)* [Text]. http://www.natcorp.ox.ac.uk/

Brunswik, E., & Kamiya, J. (1953). Ecological Cue-Validity of 'Proximity' and of Other Gestalt Factors. *The American Journal of Psychology*, *66*(1), 20–32. https://doi.org/10.2307/1417965

Chalk, M., Seitz, A. R., & Seriès, P. (2010). Rapidly learned stimulus expectations alter perception of motion. *Journal of Vision*, *10*(8), 2. https://doi.org/10.1167/10.8.2

Chang, R., Baria, A. T., Flounders, M. W., & He, B. J. (2016). Unconsciously elicited perceptual prior. *Neuroscience of Consciousness*, *2016*(1). https://doi.org/10.1093/nc/niw008

Christensen, J. H., Bex, P. J., & Fiser, J. (2015). Prior implicit knowledge shapes human threshold for orientation noise. *Journal of Vision*, *15*(9), 24–24. https://doi.org/10.1167/15.9.24

Civille, G. V., & Dus, C. A. (1990). Development of Terminology to Describe the Handfeel Properties of Paper and Fabrics. *Journal of Sensory Studies*, *5*(1), 19–32. https://doi.org/10.1111/j.1745-459X.1990.tb00479.x

Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J.-R., Muñoz-Moldes, S., Vuillaume, L., & de Heering, A. (2020). Learning to Be Conscious. *Trends in Cognitive Sciences*, *24*(2), 112–123. https://doi.org/10.1016/j.tics.2019.11.011

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8), 358–364. https://doi.org/10.1016/j.tics.2011.06.008

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002

DeGutis, J., Cohan, S., & Nakayama, K. (2014). Holistic face training enhances face processing in developmental prosopagnosia. *Brain*, *137*(6), 1781–1798. https://doi.org/10.1093/brain/awu062

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204–211. https://doi.org/10.1016/j.tics.2006.03.007

Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J.-F., Poline, J.-B., & Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, *4*(7), Article 7. https://doi.org/10.1038/89551

Del Cul, A., Baillet, S., & Dehaene, S. (2007). Brain Dynamics Underlying the Nonlinear Threshold for Access to Consciousness. *PLoS Biology*, *5*(10). https://doi.org/10.1371/journal.pbio.0050260

Derda, M., Koculak, M., Windey, B., Gociewicz, K., Wierzchoń, M., Cleeremans, A., & Binder, M. (2019). The role of levels of processing in disentangling the ERP signatures of conscious visual processing. *Consciousness and Cognition*, *73*, 102767. https://doi.org/10.1016/j.concog.2019.102767

Devezer, B., & Penders, B. (2023). Scientific reform, citation politics and the bureaucracy of oblivion. *Quantitative Science Studies*, 1–6. https://doi.org/10.1162/qss_c_00274

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In M. Overgaard (Ed.), *Behavioral Methods in Consciousness Research* (pp. 199–220). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199688890.003.0012

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Dienes, Z., & Seth, A. (2010a). 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), 674–681. https://doi.org/10.1016/j.concog.2009.09.009

Dienes, Z., & Seth, A. K. (2010b). Measuring any conscious content versus measuring the relevant conscious content: Comment on Sandberg et al. *Consciousness and Cognition*, *19*(4), 1079–1080. https://doi.org/10.1016/j.concog.2010.03.009

Doorn, J. van, Bergh, D. van den, Bohm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Raj, A., Sarafoglou, A., Stefan, A. M., Voelkel, J. G., & Wagenmakers, E.-J. (2019). *The JASP Guidelines for Conducting and Reporting a Bayesian Analysis*. PsyArXiv. https://doi.org/10.31234/osf.io/yqxfr

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., Parsonage, H., Ritchie, S., Ren, K., Tan, X., Saporta, R., Seiskari, O., Dong, X., Lang, M., Iwasaki, W., Wenchel, S., … Schwen, B. (2021). *data.table: Extension of 'data.frame'* (1.14.2) [Computer software]. https://CRAN.R-project.org/package=data.table

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170–177. https://doi.org/10.1037/1082-989X.1.2.170

Ehrenstein, W. H., & Ehrenstein, A. (1999). Psychophysical Methods. In U. Windhorst & H. Johansson (Eds.), *Modern Techniques in Neuroscience Research* (pp. 1211–1241). Springer. https://doi.org/10.1007/978-3-642-58552-4_43

Eiserbeck, A., Enge, A., Rabovsky, M., & Abdel Rahman, R. (2022). Electrophysiological Chronometry of Graded Consciousness during the Attentional Blink. *Cerebral Cortex*, *32*(6), 1244–1259. https://doi.org/10.1093/cercor/bhab289

Eklund, R., & Wiens, S. (2018). Visual awareness negativity is an early neural correlate of awareness: A preregistered study with two Gabor sizes. *Cognitive, Affective, & Behavioral Neuroscience*, *18*(1), 176–188. https://doi.org/10.3758/s13415-018-0562-z

Eriksson, J., Fontan, A., & Pedale, T. (2020). Make the Unconscious Explicit to Boost the Science of Consciousness. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00260

Faivre, N., Dubois, J., Schwartz, N., & Mudrik, L. (2019). Imaging object-scene relations processing in visible and invisible natural scenes. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-38654-z

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1), 10. https://doi.org/10.1167/7.1.10

Fine, I., & Jacobs, R. A. (2002). Comparing perceptual learning across tasks: A review. *Journal of Vision*, *2*(2), 5. https://doi.org/10.1167/2.2.5

Fiorentini, A., & Berardi, N. (1980). Perceptual learning specific for orientation and spatial frequency. *Nature*, *287*(5777), Article 5777. https://doi.org/10.1038/287043a0

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Francken, J. C., Beerendonk, L., Molenaar, D., Fahrenfort, J. J., Kiverstein, J. D., Seth, A. K., & van Gaal, S. (2022). An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. *Neuroscience of Consciousness*, *2022*(1), niac011. https://doi.org/10.1093/nc/niac011

Furmanski, C. S., & Engel, S. A. (2000). Perceptual learning in object recognition: Object specificity and size invariance. *Vision Research*, *40*(5), 473–484. https://doi.org/10.1016/S0042-6989(99)00134-0

Furmanski, C. S., Schluppeck, D., & Engel, S. A. (2004). Learning Strengthens the Response of Primary Visual Cortex to Simple Patterns. *Current Biology*, *14*(7), 573–578. https://doi.org/10.1016/j.cub.2004.03.032

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise Differences in the Comprehension of Visualizations: A Meta-Analysis of Eye-Tracking Research in Professional Domains. *Educational Psychology Review*, *23*(4), 523–552. https://doi.org/10.1007/s10648-011-9174-7

Geisler, W. S. (2008). Visual Perception and the Statistical Properties of Natural Scenes. *Annual Review of Psychology*, *59*(1), 167–192. https://doi.org/10.1146/annurev.psych.58.110405.085632

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian Regression Models. *The American Statistician*, *73*(3), 307–309. https://doi.org/10.1080/00031305.2018.1549100

Gibson, E. J. (1969). *Principles of perceptual learning and development*. Appleton-Century-Crofts.

Goldstone, R. L. (1998). Perceptual Learning. *Annual Review of Psychology*, *49*(1), 585–612. https://doi.org/10.1146/annurev.psych.49.1.585

Goldstone, R. L., & Byrge, L. A. (2015). Perceptual Learning. In M. Matthen (Ed.), *The Oxford Handbook of Philosophy of Perception* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199600472.013.029

González-García, C., Flounders, M. W., Chang, R., Baria, A. T., & He, B. J. (2018). Content-specific activity in frontoparietal and default-mode networks during prior-guided visual perception. *eLife*, *7*, e36068. https://doi.org/10.7554/eLife.36068

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence-Based Nursing*, *18*(3), 66–67. https://doi.org/10.1136/eb-2015-102129

Heeks, F., & Azzopardi, P. (2015). Thresholds for detection and awareness of masked facial stimuli. *Consciousness and Cognition*, *32*, 68–78. https://doi.org/10.1016/j.concog.2014.09.009

Heisz, J. J., & Shore, D. I. (2008). More efficient scanning for familiar faces. *Journal of Vision*, *8*(1), 9. https://doi.org/10.1167/8.1.9

Herrick, R. M. (1973). Psychophysical methodology: VI. Random method of limits. *Perception & Psychophysics*, *13*(3), 548–554. https://doi.org/10.3758/BF03205818

Hock, H., & Schöner, G. (2010). Measuring Perceptual Hysteresis with the Modified Method of Limits: Dynamics at the Threshold. *Seeing and Perceiving*, *23*(2), 173–195. https://doi.org/10.1163/187847510X503597

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences*, *9*(1), 1–23. https://doi.org/10.1017/S0140525X00021269

Hua, T., Bao, P., Huang, C.-B., Wang, Z., Xu, J., Zhou, Y., & Lu, Z.-L. (2010). Perceptual Learning Improves Contrast Sensitivity of V1 Neurons in Cats. *Current Biology*, *20*(10), 887–894. https://doi.org/10.1016/j.cub.2010.03.066

Hussain, Z., Sekuler, A. B., & Bennett, P. J. (2009). How much practice is needed to produce perceptual learning? *Vision Research*, *49*(21), 2624–2634. https://doi.org/10.1016/j.visres.2009.08.022

Irvine, E. (2012). Old Problems with New Measures in the Science of Consciousness. *The British Journal for the Philosophy of Science*, *63*(3), 627–648. https://doi.org/10.1093/bjps/axs019

Irvine, E. (2013). Measures of Consciousness. *Philosophy Compass*, *8*(3), 285–297. https://doi.org/10.1111/phc3.12016

Jannati, A., & Di Lollo, V. (2012). Relative blindsight arises from a criterion confound in metacontrast masking: Implications for theories of consciousness. *Consciousness and Cognition*, *21*(1), 307–314. https://doi.org/10.1016/j.concog.2011.10.003

Jimenez, M., Grassini, S., Montoro, P. R., Luna, D., & Koivisto, M. (2018). Neural correlates of visual awareness at stimulus low vs. High-levels of processing. *Neuropsychologia*, *121*, 144–152. https://doi.org/10.1016/j.neuropsychologia.2018.11.001

Jimenez, M., Poch, C., Villalba-García, C., Sabater, L., Hinojosa, J. A., Montoro, P. R., & Koivisto, M. (2021). The Level of Processing Modulates Visual Awareness: Evidence from Behavioral and Electrophysiological Measures. *Journal of Cognitive Neuroscience*, *33*(7), 1295–1310. https://doi.org/10.1162/jocn_a_01712

Jimenez, M., Prieto, A., Gómez, P., Hinojosa, J. A., & Montoro, P. R. (2023). Masked priming under the Bayesian microscope: Exploring the integration of local elements into global shape through Bayesian model comparison. *Consciousness and Cognition*, *115*, 103568. https://doi.org/10.1016/j.concog.2023.103568

Jimenez, M., Villalba-García, C., Luna, D., Hinojosa, J. A., & Montoro, P. R. (2019). The nature of visual awareness at stimulus energy and feature levels: A backward masking study. *Attention, Perception, & Psychophysics*, *81*(6), 1926–1943. https://doi.org/10.3758/s13414-019-01732-5

Judd, C. H., & Cowling, D. J. (1907). Studies in perceptual development. *The Psychological Review: Monograph Supplements*, *8*(3), 349–369. https://doi.org/10.1037/h0093043

Karni, A., & Sagi, D. (1993). The time course of learning a visual skill. *Nature*, *365*(6443), 250–252. https://doi.org/10.1038/365250a0

Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J. M., & Sagi, D. (1994). Dependence on REM Sleep of Overnight Improvement of a Perceptual Skill. *Science*, *265*(5172), 679–682. https://doi.org/10.1126/science.8036518

Kingdom, F. A. A., & Prins, N. (2016). *Psychophysics: A Practical Introduction*. Academic Press.

Knotts, J., Odegaard, B., Lau, H., & Rosenthal, D. (2019). Subjective inflation: Phenomenology's get-rich-quick scheme. *Current Opinion in Psychology*, *29*, 49–55. https://doi.org/10.1016/j.copsyc.2018.11.006

Koch, C., & Preuschoff, K. (2007). Betting the house on consciousness. *Nature Neuroscience*, *10*(2), 140–141. https://doi.org/10.1038/nn0207-140

Koivisto, M., Kastrati, G., & Revonsuo, A. (2013). Recurrent Processing Enhances Visual Awareness but Is Not Necessary for Fast Categorization of Natural Scenes. *Journal of Cognitive Neuroscience*, *26*(2), 223–231. https://doi.org/10.1162/jocn_a_00486

Koivisto, M., & Rientamo, E. (2016). Unconscious vision spots the animal but not the dog: Masked priming of natural scenes. *Consciousness and Cognition*, *41*, 10–23. https://doi.org/10.1016/j.concog.2016.01.008

Kok, E. M., Bruin, A. B. H. de, Robben, S. G. F., & Merriënboer, J. J. G. van. (2012). Looking in the Same Manner but Seeing it Differently: Bottom-up and Expertise Effects in Radiology. *Applied Cognitive Psychology*, *26*(6), 854–862. https://doi.org/10.1002/acp.2886

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron*, *75*(2), 265–270. https://doi.org/10.1016/j.neuron.2012.04.034

Kouider, S., de Gardelle, V., Sackur, J., & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences*, *14*(7), 301–307. https://doi.org/10.1016/j.tics.2010.04.006

Kugele, S., & Franklin, S. (2021). Learning in LIDA. *Cognitive Systems Research*, *66*, 176–200. https://doi.org/10.1016/j.cogsys.2020.11.001

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2016, January 14). The 20% Statistician: Power analysis for default Bayesian t-tests. *The 20% Statistician*. http://daniellakens.blogspot.com/2016/01/power-analysis-for-default-bayesian-t.html

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001

Lamme, V. A. F. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience*, *1*(3), 204–220. https://doi.org/10.1080/17588921003731586

Lamme, V. A. F. (2014). The Crack of Dawn: Perceptual Functions and Neural Mechanisms that Mark the Transition from Unconscious Processing to Conscious Vision. In T. Metzinger & J. M. Windt (Eds.), *Open MIND*. Open MIND. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570092

Lamy, D., Alon, L., Carmel, T., & Shalev, N. (2015). The Role of Conscious Perception in Attentional Capture and Object-File Updating. *Psychological Science*, *26*(1), 48–57. https://doi.org/10.1177/0956797614556777

Lau, H. C. (2007). A higher order Bayesian decision theory of consciousness. In R. Banerjee & B. K. Chakrabarti (Eds.), *Progress in Brain Research* (Vol. 168, pp. 35–48). Elsevier. https://doi.org/10.1016/S0079-6123(07)68004-2

Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences*, *103*(49), 18763–18768. https://doi.org/10.1073/pnas.0607716103

Liang, J. C., Erez, J., Zhang, F., Cusack, R., & Barense, M. D. (2020). Experience Transforms Conjunctive Object Representations: Neural Evidence for Unitization After Visual Expertise. *Cerebral Cortex*, *30*(5), 2721–2739. https://doi.org/10.1093/cercor/bhz250

Lin, Z., Dosher, B. A., & Lu, Z.-L. (2017). Mixture of easy trials enables transient and sustained perceptual improvements through priming and perceptual learning. *Scientific Reports*, *7*(1), 7421. https://doi.org/10.1038/s41598-017-06989-0

Liu, J., Lu, Z.-L., & Dosher, B. A. (2012). Mixed training at high and low accuracy levels leads to perceptual learning without feedback. *Vision Research*, *61*, 15–24. https://doi.org/10.1016/j.visres.2011.12.002

Lu, Z.-L., & Dosher, B. A. (2022). Current directions in visual perceptual learning. *Nature Reviews Psychology*, *1*(11), Article 11. https://doi.org/10.1038/s44159-022-00107-2

Ludmer, R., Dudai, Y., & Rubin, N. (2011). Uncovering camouflage: Amygdala activation predicts long-term memory of induced perceptual insight. *Neuron*, *69*(5), 1002–1014. https://doi.org/10.1016/j.neuron.2011.02.013

Lupyan, G. (2017). Objective effects of knowledge on visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(4), 794–806. https://doi.org/10.1037/xhp0000343

Makowski, D. (2018). The psycho Package: An Efficient and Publishing-Oriented Workflow for Psychological Science. *Journal of Open Source Software*, *3*(22), 470. https://doi.org/10.21105/joss.00470

Markant, J., & Scott, L. S. (2018). Attention and Perceptual Learning Interact in the Development of the Other-Race Effect. *Current Directions in Psychological Science*, *27*(3), 163–169. https://doi.org/10.1177/0963721418769884

MATLAB (9.10.0.1602886 (R2021a)). (2021). [Computer software]. The MathWorks Inc.

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual Skills in Airport-Security Screening. *Psychological Science*, *15*(5), 302–306. https://doi.org/10.1111/j.0956-7976.2004.00673.x

McKee, S. P., & Westheimer, G. (1978). Improvement in vernier acuity with practice. *Perception & Psychophysics*, *24*(3), 258–262. https://doi.org/10.3758/BF03206097

Mednick, S. C., Drummond, S. P. A., Boynton, G. M., Awh, E., & Serences, J. (2008). Sleep-Dependent Learning and Practice-Dependent Deterioration in an Orientation Discrimination Task. *Behavioral Neuroscience*, *122*(2), 267–272. https://doi.org/10.1037/0735-7044.122.2.267

Merikle, P. M., Smilek, D., & Eastwood, J. D. (2001). Perception without awareness: Perspectives from cognitive psychology. *Cognition*, *79*(1), 115–134. https://doi.org/10.1016/S0010-0277(00)00126-8

Metcalfe, J. (1986). Premonitions of insight predict impending error. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(4), 623–634. https://doi.org/10.1037/0278-7393.12.4.623

Meuwese, J. D. I., Post, R. A. G., Scholte, H. S., & Lamme, V. A. F. (2013). Does Perceptual Learning Require Consciousness or Attention? *Journal of Cognitive Neuroscience*, *25*(10), 1579–1596. https://doi.org/10.1162/jocn_a_00424

Meyen, S., Zerweck, I. A., Amado, C., von Luxburg, U., & Franz, V. H. (2022). Advancing research on unconscious priming: When can scientists claim an indirect task advantage? *Journal of Experimental Psychology: General*, *151*(1), 65–81. https://doi.org/10.1037/xge0001065

Michel, M. (2019). The Mismeasure of Consciousness: A Problem of Coordination for the Perceptual Awareness Scale. *Philosophy of Science*, *86*(5), 1239–1249. https://doi.org/10.1086/705509

Michel, M. (2022). How (not) to underestimate unconscious perception. *Mind & Language*, *n/a*(n/a), 1–18. https://doi.org/10.1111/mila.12406

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2022). *BayesFactor: Computation of Bayes Factors for Common Designs* (0.9.12-4.4) [Computer software]. https://CRAN.R-project.org/package=BayesFactor

Müller, K., & Bryan, J. (2020). *here: A Simpler Way to Find Your Files* (1.0.1) [Computer software]. https://CRAN.R-project.org/package=here

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435. https://doi.org/10.2307/2183914

Neri, P. (2014). Semantic Control of Feature Extraction from Natural Scenes. *Journal of Neuroscience*, *34*(6), 2374–2388. https://doi.org/10.1523/JNEUROSCI.1755-13.2014

Nishina, S., Seitz, A. R., Kawato, M., & Watanabe, T. (2007). Effect of spatial distance to the task stimulus on task-irrelevant perceptual learning of static Gabors. *Journal of Vision*, *7*(13), 2. https://doi.org/10.1167/7.13.2

Norman, E., & Price, M. (2015). Measuring consciousness with confidence ratings. In M. Overgaard (Ed.), *Behavioural Methods in Consciousness Research* (pp. 159–180). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199688890.003.0010

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Overgaard, M. (2018). Phenomenal consciousness and cognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170353. https://doi.org/10.1098/rstb.2017.0353

Overgaard, M., Feldbæk Nielsen, J., & Fuglsang-Frederiksen, A. (2004). A TMS study of the ventral projections from V1 with implications for the finding of neural correlates of consciousness. *Brain and Cognition*, *54*(1), 58–64. https://doi.org/10.1016/S0278-2626(03)00260-4

Overgaard, M., Rote, J., Mouridsen, K., & Ramsøy, T. Z. (2006). Is conscious perception gradual or dichotomous? A comparison of report methodologies during a visual task.

Consciousness and Cognition, *15*(4), 700–708. https://doi.org/10.1016/j.concog.2006.04.002

Overgaard, M., & Sandberg, K. (2012). Kinds of access: Different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1287–1296. https://doi.org/10.1098/rstb.2011.0425

Overgaard, M., & Sandberg, K. (2021). The Perceptual Awareness Scale—Recent controversies and debates. *Neuroscience of Consciousness*, *2021*(1), niab044. https://doi.org/10.1093/nc/niab044

Pascucci, D., Mastropasqua, T., & Turatto, M. (2015). Monetary Reward Modulates Task-Irrelevant Perceptual Learning for Invisible Stimuli. *PLOS ONE*, *10*(5), e0124009. https://doi.org/10.1371/journal.pone.0124009

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Peremen, Z., & Lamy, D. (2014). Comparing unconscious processing during continuous flash suppression and meta-contrast masking just under the limen of consciousness. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00969

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, *10*(2), Article 2. https://doi.org/10.1038/nn1840

Pessiglione, M., Petrovic, P., Daunizeau, J., Palminteri, S., Dolan, R. J., & Frith, C. D. (2008). Subliminal Instrumental Conditioning Demonstrated in the Human Brain. *Neuron*, *59*(4), 561–567. https://doi.org/10.1016/j.neuron.2008.07.005

Peters, M. A. K., Kentridge, R. W., Phillips, I., & Block, N. (2017). Does unconscious perception really exist? Continuing the ASSC20 debate. *Neuroscience of Consciousness*, *2017*(nix015). https://doi.org/10.1093/nc/nix015

Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *eLife*, *4*, e09651. https://doi.org/10.7554/eLife.09651

Poggio, T., Fahle, M., & Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science*, *256*(5059), 1018–1021. https://doi.org/10.1126/science.1589770

Pollack, I. (1968). Computer Simulation of Threshold Observations by Method of Limits. *Perceptual and Motor Skills*, *26*(2), 583–586. https://doi.org/10.2466/pms.1968.26.2.583

Prettyman, A. (2019). Perceptual learning. *WIREs Cognitive Science*, *10*(3), e1489. https://doi.org/10.1002/wcs.1489

Prins, N., & Kingdom, F. A. A. (2018). Applying the Model-Comparison Approach to Test Specific Research Hypotheses in Psychophysical Research Using the Palamedes Toolbox. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/article/10.3389/fpsyg.2018.01250

Prochazkova, E., Venneker, D., De Zwart, R., Tamietto, M., & Kret, M. E. (2022). Conscious awareness is necessary to assess trust and mimic facial expressions, while pupils impact trust unconsciously. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1863), 20210183. https://doi.org/10.1098/rstb.2021.0183

Quinn, P. C., & Bhatt, R. S. (2005). Learning Perceptual Organization in Infancy. *Psychological Science*, *16*(7), 511–515. https://doi.org/10.1111/j.0956-7976.2005.01567.x

Quinn, P. C., & Bhatt, R. S. (2006). Are some Gestalt principles deployed more readily than others during early development? The case of lightness versus form similarity. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1221–1230. https://doi.org/10.1037/0096-1523.32.5.1221

Quintana, D. S., & Williams, D. R. (2018). Bayesian alternatives for common null-hypothesis significance tests in psychiatry: A non-technical guide using JASP. *BMC Psychiatry*, *18*(1), 178. https://doi.org/10.1186/s12888-018-1761-4

R Core Team. (2021). *R: A language and environment for statistical computing* (2021.09.1+372 'Ghost Orchid' Release) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rajananda, S., Zhu, J., & Peters, M. A. K. (2020). Normal observers show no evidence for blindsight in facial emotion perception. *Neuroscience of Consciousness*, *2020*(niaa023). https://doi.org/10.1093/nc/niaa023

Ramamurthy, U., Baars, B. J., S. K., D., & Franklin, S. (2006). LIDA: A Working Model of Cognition. *Proceedings of the 7th International Conference on Cognitive Modeling*, (pp. 244-249). https://web-archive.southampton.ac.uk/cogprints.org/5852/

Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences*, *3*(1), 1–23. https://doi.org/10.1023/B:PHEN.0000041900.30172.e8

Ransom, M. (2020). Attentional Weighting in Perceptual Learning. *Journal of Consciousness Studies*, *27*(7–8), 236–248.

Reingold, E. M., & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics*, *44*(6), 563–575. https://doi.org/10.3758/BF03207490

Rohatgi, A. (2015). *WebPlotDigitizer* (3.9) [Computer software]. https://apps.automeris.io/wpd/

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Rousselet, G., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, *12*(6), 852–877. https://doi.org/10.1080/13506280444000553

RStudio Team. (2021). *RStudio: Integrated Development for R.* [Computer software]. RStudio, PBC. http://www.rstudio.com/

Sagi, D. (2011). Perceptual learning in Vision Research. *Vision Research*, *51*(13), 1552–1566. https://doi.org/10.1016/j.visres.2010.10.019

Samaha, J., Boutonnet, B., Postle, B. R., & Lupyan, G. (2018). Effects of meaningfulness on perception: Alpha-band oscillations carry perceptual expectations and influence early visual responses. *Scientific Reports*, *8*(1), Article 1. https://doi.org/10.1038/s41598-018-25093-5

Sand, A., & Nilsson, M. E. (2017). When Perception Trumps Reality: Perceived, Not Objective, Meaning of Primes Drives Stroop Priming. *Psychological Science*, *28*(3), 346–355. https://doi.org/10.1177/0956797616684681

Sandberg, K., Bibby, B. M., Timmermans, B., Cleeremans, A., & Overgaard, M. (2011). Measuring consciousness: Task accuracy and awareness as sigmoid functions of

stimulus duration. *Consciousness and Cognition*, *20*(4), 1659–1675. https://doi.org/10.1016/j.concog.2011.09.002

Sandberg, K., & Overgaard, M. (2015). Using the perceptual awareness scale (PAS). In M. Overgaard (Ed.), *Behavioral Methods in Consciousness Research* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199688890.003.0011

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*, *19*(4), 1069–1078. https://doi.org/10.1016/j.concog.2009.12.013

Sanders, M. D., Warrington, E. K., Marshall, J., & Wieskrantz, L. (1974). 'Blindsight': Vision in a field defect. *The Lancet*, *303*(7860), 707–708. https://doi.org/10.1016/S0140-6736(74)92907-9

Schlaghecken, F., Blagrove, E., & Maylor, E. A. (2008). No difference between conscious and nonconscious visuomotor control: Evidence from perceptual learning in the masked prime task. *Consciousness and Cognition*, *17*(1), 84–93. https://doi.org/10.1016/j.concog.2006.11.004

Scholes, C., McGraw, P. V., & Roach, N. W. (2021). Learning to silence saccadic suppression. *Proceedings of the National Academy of Sciences*, *118*(6), e2012937118. https://doi.org/10.1073/pnas.2012937118

Schwiedrzik, C. M., Melloni, L., & Schurger, A. (2018). Mooney face stimuli for visual perception research. *PLoS ONE*, *13*(7). https://doi.org/10.1371/journal.pone.0200106

Schwiedrzik, C. M., Singer, W., & Melloni, L. (2009). Sensitivity and perceptual awareness increase with practice in metacontrast masking. *Journal of Vision*, *9*(10), 18–18. https://doi.org/10.1167/9.10.18

Schwiedrzik, C. M., Singer, W., & Melloni, L. (2011). Subjective and objective learning effects dissociate in space and in time. *Proceedings of the National Academy of Sciences*, *108*(11), 4506–4511. https://doi.org/10.1073/pnas.1009147108

Schyns, P. G., & Oliva, A. (1994). From Blobs to Boundary Edges: Evidence for Time- and Spatial-Scale-Dependent Scene Recognition. *Psychological Science*, *5*(4), 195–200. https://doi.org/10.1111/j.1467-9280.1994.tb00500.x

Seitz, A. R., Kim, D., & Watanabe, T. (2009). Rewards Evoke Learning of Unconsciously Processed Visual Stimuli in Adult Humans. *Neuron*, *61*(5), 700–707. https://doi.org/10.1016/j.neuron.2009.01.016

Seitz, A. R., Lefebvre, C., Watanabe, T., & Jolicoeur, P. (2005). Requirement for high-level processing in subliminal learning. *Current Biology*, *15*(18), R753–R755.

Sergent, C., & Dehaene, S. (2004). Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science*, *15*(11), 720–728. https://doi.org/10.1111/j.0956-7976.2004.00748.x

Series, P., & Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00668

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, 1–14. https://doi.org/10.1038/s41583-022-00587-4

Sidis, B. (1898). *The Psychology of Suggestion: A Research Into the Subconscious Nature of Man and Society*. D. Appleton.

Siedlecka, M., Wereszczyński, M., Paulewicz, B., & Wierzchoń, M. (2020). Visual awareness judgments are sensitive to accuracy feedback in stimulus discrimination tasks. *Consciousness and Cognition*, *86*, 103035. https://doi.org/10.1016/j.concog.2020.103035

Skewes, J., Frith, C., & Overgaard, M. (2021). Awareness and confidence in perceptual decision-making. *Brain Multiphysics*, *2*, 100030. https://doi.org/10.1016/j.brain.2021.100030

Skora, L. I., & Scott, R. B. (2022). *Stimulus awareness is necessary for both instrumental learning and instrumental responding to previously learned stimuli*. PsyArXiv. https://doi.org/10.31234/osf.io/e6sw8

Skora, L. I., Yeomans, M. R., Crombag, H. S., & Scott, R. B. (2021). Evidence that instrumental conditioning requires conscious awareness in humans. *Cognition*, *208*, 104546. https://doi.org/10.1016/j.cognition.2020.104546

Skóra, Z., Del Pin, S. H., Derda, M., Koculak, M., Rutiku, R., & Wierzchoń, M. (2021). No validity without a theory-a critical look at subjective measures of consciousness. *Neuroscience of Consciousness*, *2021*(1), niab009. https://doi.org/10.1093/nc/niab009

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*(6), 2083–2101. https://doi.org/10.3758/s13423-018-1451-8

Sowden, P. T., Davies, I. R. L., & Roling, P. (2000). Perceptual learning of the detection of features in X-ray images: A functional role for improvements in adults' visual sensitivity? *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 379–390. https://doi.org/10.1037/0096-1523.26.1.379

Sowden, P. T., Rose, D., & Davies, I. R. L. (2002). Perceptual learning of luminance contrast detection: Specific for spatial frequency and retinal location but not orientation. *Vision Research*, *42*(10), 1249–1258. https://doi.org/10.1016/S0042-6989(02)00019-6

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11), 1–29. https://doi.org/10.1037/h0093759

Stein, T., & Peelen, M. V. (2021). Dissociating conscious and unconscious influences on visual detection effects. *Nature Human Behaviour*, *5*(5), 612–624. https://doi.org/10.1038/s41562-020-01004-5

Stein, T., Utz, V., & van Opstal, F. (2020). Unconscious semantic priming from pictures under backward masking and continuous flash suppression. *Consciousness and Cognition*, *78*, 102864. https://doi.org/10.1016/j.concog.2019.102864

Stickgold, R., Whidbee, D., Schirmer, B., Patel, V., & Hobson, J. A. (2000). Visual Discrimination Task Improvement: A Multi-Step Process Occurring During Sleep. *Journal of Cognitive Neuroscience*, *12*(2), 246–254. https://doi.org/10.1162/089892900562075

Strasburger, H. (2001). Converting between measures of slope of the psychometric function. *Perception & Psychophysics*, *63*(8), 1348–1355. https://doi.org/10.3758/BF03194547

Szczepanowski, R., Traczyk, J., Wierzchoń, M., & Cleeremans, A. (2013). The perception of visual emotion: Comparing different measures of awareness. *Consciousness and Cognition*, *22*(1), 212–220. https://doi.org/10.1016/j.concog.2012.12.003

Tendeiro, J. N., Kiers, H. A. L., Hoekstra, R., Wong, T. K., & Morey, R. D. (2024). Diagnosing the Misuse of the Bayes Factor in Applied Research. *Advances in Methods and Practices in Psychological Science*, *7*(1), 25152459231213371. https://doi.org/10.1177/25152459231213371

Teufel, C., Dakin, S. C., & Fletcher, P. C. (2018). Prior object-knowledge sharpens properties of early visual feature-detectors. *Scientific Reports*, *8*(1), 10853. https://doi.org/10.1038/s41598-018-28845-5

Teufel, C., & Fletcher, P. C. (2020). Forms of prediction in the nervous system. *Nature Reviews Neuroscience*, *21*(4), 231–242. https://doi.org/10.1038/s41583-020-0275-5

Teufel, C., & Nanay, B. (2017). How to (and how not to) think about top-down influences on visual perception. *Consciousness and Cognition*, *47*, 17–25. https://doi.org/10.1016/j.concog.2016.05.008

Teufel, C., Subramaniam, N., Dobler, V., Perez, J., Finnemann, J., Mehta, P. R., Goodyer, I. M., & Fletcher, P. C. (2015). Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proceedings of the National Academy of Sciences*, *112*(43), 13401–13406. https://doi.org/10.1073/pnas.1503916112

Thiruvasagam, S., & Srinivasan, N. (2021). Gradedness of visual awareness depends on attentional scope: Global perception is more graded than local perception. *Consciousness and Cognition*, *94*, 103174. https://doi.org/10.1016/j.concog.2021.103174

Thiruvasagam, S., & Srinivasan, N. (2023). Corrigendum to “Gradedness of visual awareness depends on attentional scope: Global perception is more graded than local perception” [Conscious. Cogn. 94 (2021) 103174]. *Consciousness and Cognition*, *109*, 103489. https://doi.org/10.1016/j.concog.2023.103489

Tiedemann, F. (2022). *gghalves: Compose Half-Half Plots Using Your Favourite Geoms* (0.1.4) [Computer software]. https://CRAN.R-project.org/package=gghalves

Timmermans, B., & Cleeremans, A. (2015). How can we measure awareness? An overview of current methods. In *Behavioral Methods in Consciousness Research* (pp. 21–46). Oxford University Press. https://abdn.pure.elsevier.com/en/publications/how-can-we-measure-awareness-an-overview-of-current-methods

Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in action: Consciousness as an unconscious re-description process. *Philosophical*

Transactions of the Royal Society B: Biological Sciences, *367*(1594), 1412–1423. https://doi.org/10.1098/rstb.2011.0421

Travers, E., Frith, C. D., & Shea, N. (2018). Learning rapidly about the relevance of visual cues requires conscious awareness. *Quarterly Journal of Experimental Psychology*, *71*(8), 1698–1713. https://doi.org/10.1080/17470218.2017.1373834

Tsuchiya, N., & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. *Nature Neuroscience*, *8*(8), Article 8. https://doi.org/10.1038/nn1500

Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. F. (2015). No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. *Trends in Cognitive Sciences*, *19*(12), 757–770. https://doi.org/10.1016/j.tics.2015.10.002

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*, *23*(1), 87–102. https://doi.org/10.3758/s13423-015-0892-6

Van den Bussche, E., Notebaert, K., & Reynvoet, B. (2009). Masked primes can be genuinely semantically processed: A picture prime study. *Experimental Psychology*, *56*(5), 295–300. https://doi.org/10.1027/1618-3169.56.5.295

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & Heydt, R. von der. (2012). A Century of Gestalt Psychology in Visual Perception I. Perceptual Grouping and Figure-Ground Organization. *Psychological Bulletin*, *138*(6), 1172. https://doi.org/10.1037/a0029333

Walker, M. P., & Stickgold, R. (2004). Sleep-Dependent Learning and Memory Consolidation. *Neuron*, *44*(1), 121–133. https://doi.org/10.1016/j.neuron.2004.08.031

Watanabe, T., Náñez, J. E., & Sasaki, Y. (2001). Perceptual learning without perception. *Nature*, *413*(6858), 844–848. https://doi.org/10.1038/35101601

Watanabe, T., & Sasaki, Y. (2015). Perceptual Learning: Toward a Comprehensive Theory. *Annual Review of Psychology*, *66*(1), 197–221. https://doi.org/10.1146/annurev-psych-010814-015214

Watson, A. B. (2017). QUEST+: A general multidimensional Bayesian adaptive psychometric method. *Journal of Vision*, *17*(3), 10. https://doi.org/10.1167/17.3.10

Westfall, J. (2016, March 25). Five different "Cohen's d" statistics for within-subject designs. *Cookie Scientist*. http://jakewestfall.org/blog/index.php/2016/03/25/five-different-cohens-d-statistics-for-within-subject-designs/

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science*, *6*(3), 291–298. https://doi.org/10.1177/1745691611406923

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Bryan, J. (2023). *readxl: Read Excel Files* (1.4.2) [Computer software]. https://CRAN.R-project.org/package=readxl

Wickham, H., Pedersen, T. L., Seidel, D., Posit, & PBC. (2023). *scales: Scale Functions for Visualization* (1.2.0) [Computer software]. https://cloud.r-project.org/web/packages/scales/index.html

Wickham, H., & RStudio. (2023). *tidyverse: Easily Install and Load the 'Tidyverse'* (2.0.0) [Computer software]. https://cran.r-project.org/web/packages/tidyverse/index.html

Wierzchoń, M., Anzulewicz, A., Hobot, J., Paulewicz, B., & Sackur, J. (2019). In search of the optimal measure of awareness: Discrete or continuous? *Consciousness and Cognition*, *75*, 102798. https://doi.org/10.1016/j.concog.2019.102798

Wierzchoń, M., Asanowicz, D., Paulewicz, B., & Cleeremans, A. (2012). Subjective measures of consciousness in artificial grammar learning task. *Consciousness and Cognition*, *21*(3), 1141–1153. https://doi.org/10.1016/j.concog.2012.05.012

Wierzchoń, M., Paulewicz, B., Asanowicz, D., Timmermans, B., & Cleeremans, A. (2014). Different subjective awareness measures demonstrate the influence of visual identification on perceptual awareness ratings. *Consciousness and Cognition*, *27*, 109–120. https://doi.org/10.1016/j.concog.2014.04.009

Yu, Q., Zhang, P., Qiu, J., & Fang, F. (2016). Perceptual Learning of Contrast Detection in the Human Lateral Geniculate Nucleus. *Current Biology*, *26*(23), 3176–3182. https://doi.org/10.1016/j.cub.2016.09.034

Zher-Wen, & Yu, R. (2023). Unconscious integration: Current evidence for integrative processing under subliminal conditions. *British Journal of Psychology*, *114*(2), 430–456. https://doi.org/10.1111/bjop.12631