Review by Jeffrey Saunders, 05 Dec 2022 12:04

This is a well-thought study that will contribute to the literature. It is closely based on a previous study, so the general idea is not novel. However, it will provide a better controlled test of the previously reported effect.

The authors do a good analysis of previous studies, and present a compelling case for revisiting the findings of Chang et al (2016). They discuss theoretical and empirical reasons to doubt that masked information could allow future disambiguation of two-tone images, and identify limitations of the methods of Chang et al (2016). The background and motivation for the study are clearly presented, with good arguments.

1.1) The authors have also given careful thought to the methodology. The primary challenge is ensuring that "unconscious" stimuli are truly unconscious. I think the authors do a good job meeting this challenge. Trials will be classified based on multiple measures in a graded manner. The criteria for "fully unconscious" is more conservative than in the previous study, so we can more confident that any post-exposure effects will not be due to some conscious awareness. I think this is the main feature of the new study. They have also given good consideration to details like attention checks, exclusion criteria, and statistical power. The fact that it is pre-registered RR is positive feature in itself.

We would like to thank the reviewer for the supportive feedback on our manuscript.

1.2) I question whether Experiment 2 is needed. Experiment 1 already implements blind ratings, and specifies a coding plan. Any errors or biases in rating responses would just add noise or shift the baselines. As the authors point out, adopting the forced choice method also has drawbacks, which might end up increasing the variability. More data is always nice, so if the authors want to repeat with this variation in method, that is fine. It seems like a lot of extra data collection to address an issue that is unlikely to affect the results, and might introduce some new problems.

If Experiment 2 is going to be included, the authors should say more about what they would conclude if the results from the two experiments are not entirely consistent. What if Experiment 1 finds strong evidence for unconscious priming but Experiment 2 finds only a weak trend? Would they conclude that there was experimenter bias in Experiment 1, and that the effect may not be reliable? Or conclude that the data in Experiment 2 was noisier due to methodological issues, so it should be discounted?

We consider Experiment 2 an important addition to the study's effort to test the robustness of the effect suggested by Chang and colleagues across different methodological choices. Furthermore, the first wave of data for Experiment 2 has already been collected (but not accessed, in accordance with Level 3 guidelines), and therefore it would not be desirable to disregard this data. We mentioned that data collection began on 23rd Nov 2022 in the PCI RR submission form, but not in the manuscript. Nevertheless, we appreciate that more details about the aims and prospective results of Experiment 2 would be beneficial, so we have revised the section of Experiment 2 (page 18) to include these details:

NB: following the suggestion of Reviewer #2 (point 2.13 below), we have renamed the "Main" trials as "Test", and renamed our consciousness level categories into "unconscious" (point 1.9 below) etc.

"While the two methods of collecting accuracy are distinct, we expect the overall pattern of results between Experiments 1 and 2 to be the same – i.e., if we observe a differential effect between test and catch images that were unconscious (Level U) in Experiment 1, we expect the same result in Experiment 2. Having a secondary method of testing accuracy allows us to test the robustness of the effect. Nevertheless, should we observe an effect in one experiment but not the other, this pattern 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."

1.3) I think that the third experiment makes more sense as a follow-up because it addresses a potential alternate explanation that is more likely and problematic, and which might be ruled out by the Experiment 1 results. If the main trials and catch trials don't show a difference, then the follow-up experiment will be important, but if there is a clear difference between main trials and catch trials, then it isn't needed. This is more important than the issue of subjective ratings. In fact, I suggest reversing the order. If the evidence suggests that effects in Experiment 1 are due to spontaneous disambiguation, then it would be better to know this before conducting the proposed Experiment 2, which would otherwise have the same confound.

We are not entirely sure of what the reviewer means here, but we will try to clarify.

First, the *main* effect that we are testing for in Experiment 1 (a difference between Catch and Test trials) cannot be *due to spontaneous disambiguation* (SD), because the same levels of SD are expected in both conditions. SD is therefore not a problematic confound for the effect we are testing, but rather an unavoidable potential consequence of presenting the two-tones again in the Post exposure stage. It can be one factor contributing to a Pre-Post change in performance, but one that this article is not interested in. Relatedly, the aim of Experiment 3 is not to test for the presence of SD, but instead to check if any difference between Pre and Post can be explained by anything else but SD, i.e., semantic/visual category information.

Secondly, we are unsure of the benefit of running Experiment 3 before Experiment 2. Even if we did and found evidence for SD following Experiments 1 and 3, there is no design change that we could apply to Experiment 2 to remove SD. In light of this, and prompted by the reviewer's comment, we decided it would be more informative for Experiment 3 to use the design of whichever experiment prompted the follow-up, i.e., if only Experiment 2 indicated follow-up is needed, then Experiment 3 will use the same design (MCQ). If both experiments indicate follow-up is needed, then we will only use the design of Experiment 1 (free naming). We have now revised the text to clarify this (pages 20 and 21).

1.4) For the third experiment (the results contingent follow-up), the authors should say something about the conclusions that would be drawn from different possible outcomes. What if Experiment 1 appears to show disambiguation from unconscious stimuli, but the follow-up study does not?

Thank you for highlighting this aspect. Given that the test trials in the follow-up study would only constitute a direct replication for the catch trials in Experiment 1 (or 2), should we not observe a similar Pre-Post pattern then we would interpret this as a failure to replicate the effect, and call into question the robustness of the effect. We added a paragraph on page 21 to detail some alternative interpretations:

"Experiment 3 will allow us to better interpret the findings from Experiment 1 (or 2, if the design of Experiment 2 is used). Firstly, we expect to find the same pattern of Pre-Post changes in the catch trials in Experiment 1 (or 2) and the "Related catch" trials in Experiment 3, across all categories, since the experimental conditions would be virtually identical. Should this not be observed, then the replicability of the disambiguation effect would be strongly called into question. Furthermore, if we do observe a difference between Related and Blank catch trials in Experiment 3, we would interpret this as evidence that semantic information or the category-specific low-level visual information contributed to the Pre-Post disambiguation observed in Experiment 1 (or 2). Alternatively, if we find evidence against a difference between the related and blank catch trials in Experiment 3, this will be interpreted as

evidence that the Pre-Post effect in Experiment 1 (or 2, if present) was fully driven by spontaneous disambiguation, while unconscious semantic and category-specific visual information played no role."

1.5) To evaluate the planned analysis and presentation of the results, I would like to see some sort of draft of the results section. The authors could use simulated data or placeholders for statistical results. The authors describe the planned analyses in the study design table, but there are a lot of hypotheses and analyses, and it is a bit hard to follow. Presenting the planned analyses in the format of a results section will make it easier to check that the analyses make sense and nothing is missing, and also provides an opportunity for reviewers to give feedback about the presentation.

I am not a fan of the "study design table" required by PCI-RR. Answering all the questions in a single row for each hypothesis requires a table that spans multiple pages, with narrow text blocks. The sampling plan is generally the same for all hypotheses, so that column has redundant information. The space limitation encourages enumeration of hypotheses, so a reader has to keep track of many non-descriptive labels (H1a, H1b, …). Given the limitations of the format, I think the authors did a reasonable job conveying the information. I hope that PCI-RR changes this requirement, or allows some flexibility in how the information is organized. In the meantime, it would be helpful to see the analysis plan presented as a results section.

A draft of the results section has now been added for both proposed Experiments 1 and 2, with placeholders in lieu of demographic information, trial exclusions, and statistical tests. As shown, we will include all test results in tables, for ease of comparing conditions and keeping track of all hypotheses. The results-contingent tests that will determine if Experiment 3 will be ran will be described within the paragraph rather than the table, to avoid confusion between required and optional analyses.

1.6) Using the Bayesian sequential sampling procedure is a good idea, and the proposed stopping criteria should provide good power for a range of possible effects. I have some suggestions.

For computation of Bayes' factors, the authors propose using a Cachy prior with scale parameter r = 1/sqrt(2). Schönbrodt & Wagenmakers (2018), following Rouder et al (2009), recommend a scale parameter of r = 1. They note that smaller scale parameters take longer to reach the H0 criteria in the null case. Their simulations of stopping criteria BF>6 also found that the Type I rate is slightly inflated with r = 1/sqrt(2), but not with r = 1. I suggest that they follow Schönbrodt & Wagenmakers (2018) and use r = 1.

We would like to thank the reviewer for the suggestion. In summary, we have changed the scale to r = 1. We provide details below of the justification, based on the simulations we ran.

We initially favoured a medium scale width (r = 2/sqrt(2), or 0.707) because its interpretation (i.e., 50% confidence that the true effect size (ES) lies between d of -0.707 and 0.707) appeared more plausible to us than the range of -1 to 1 that r=1 would entail (as also argued by Quintana & Williams, 2018), and was the default scale in the BayesFactor package we used. However, having ran simulations on a range of ESs (including 0 and various estimates derived from the Chang et al.'s paper) and scales of 0.707 or 1 (data and code available on OSF project page/effectSizeEstimation/effectSizeSimulations) we conclude that a change from 0.707 to 1 is reasonable. To summarise the results of the simulations with one-sided tests, in case of a true ES of 0, the change was beneficial overall. In case of a true positive but weak (0.14 to 0.42, see answer 1.7 below) or medium (0.5) ES, the change was virtually inconsequential because the false negative rate was less than 1% and the true positive only decreased by a few percentage points.

1.7) I also think that the authors should provide a justification for the choice of boundary criteria based on expected effect size, and describe the power for one or more possible effect sizes. The methods section includes statement about the boundary criteria: "A BF of 6 (or 1/6), taken to indicate moderate evidence (Lee & Wagenmakers, 2014, as cited in Quintana & Williams, 2018), was chosen as an estimated equivalent for a medium effect size." That helps make a connection from the BF criteria to effect size, but does not say anything about why a medium effect size is targeted. Later, the authors report estimated effect sizes from the previous study, but that is not connected to the choice of stopping criteria.

The boundary criteria and prior determine the range of possible effect sizes that could be reliably detected, so a given BF criteria implies a target effect size. For example, the simulation results of Schönbrodt & Wagenmakers (2018) found that a criteria of BF>6 and r=1 would have 86% power for d = 0.4 in a between-subjects, so this criteria would correspond to targeting an effect size of d ≥ 0.4. In the present study, using BF>6 will allow detection of smaller effects because it is a within-subjects design. Reporting the minimum effect size that could be reliably detected will make it easy for the reader to see that the study is well-powered (even if not familiar with BFs).

As a result of the simulations above, we can confirm that our design can detect ESs of 0.5 in 86.01% of cases and an ES of 0 in 80.2% of cases (r = 1, n = 120). An ES of 0.5 is only a default guess, as other attempts to infer an ES between Grey and Catch conditions based on the information available in Chang et al. (2016) did not prove reliable, as we explain below.

Calculating Cohen's d or a similar ES metric for a within-subject design is difficult and open to substantial variability. For example, using 5 different approaches to calculate a d-like ES, Jake Westfall demonstrated that the value can vary considerably (in their dataset, it varied between 0.25 and 1.91, (http://jakewestfall.org/blog/index.php/2016/03/25/five-different-cohens-d-statistics-for-within-subject-designs/). It is made even more difficult because the raw data is not available and all that can be accessed is the t-test statistic for the key comparison from the Chang et al.'s article (change in Pre-Post for Grey Not Recognized – labelled as unconscious test trials – versus all Catch trials). Using this (page 7, t = 2.076, p = 0.0492, n= 24, which they obtained after removing one outlier for being 3SDs outside the group mean), we obtained two different ESs based on two different methods: d_z = 0.42 (Cohen, 1988; Lakens, 2013) and d_t = 0.6 through naïve conversion from the t statistic (Dunlap et al., 1996). However, we note that the second estimate could be inflated due to correlations within the variables as a result of individual differences in spontaneous disambiguation. Using mean and SDs instead (NB: n = 25 so data is not identical to that entered to generate the t-statistic above) for catch trials extracted from Figure 2 to compute the classical Cohen's d formula – which ignores whether the data comes from between or within-subjects designs, therefore underestimating the true ES – we obtained d= 0.14, which would only be detected in 11.69% of cases even with 300 participants. From these estimates, d_t and d have clear limitations of over/underestimating the true ES; d_z might not have these limitations, but it can still differ substantially from other estimates that we could not compute because they would have required access to the data.

Because these values are conflicting, rather than picking one over the others, we revert to using a default value of 0.5, thus in the range of values computed directly from the key comparison t statistic. We have now clarified in the paper what the minimum reliably detectable ES is, and added a more detailed justification (page 13).

1.8) The lower bound on sample size, N=60, seems higher than necessary. Sequential procedures are more efficient because they can stop early if the evidence shows clear evidence one way or the other. This efficiency is lost if the lower bound is higher than needed. An effect size of d = 0.5 only needs

N=44 for 90% power. In the case of no effect, N=30 would be enough for reasonably sized confidence intervals around zero in the not-recognized condition (SE = 5.3%/sqrt(30) = 1.02%). I suggest that the authors use a smaller lower bound, N=30-40, so they can take advantage of the efficiency of the sequential testing. The sample size will still go past N=60 if the data is ambiguous, but not if the true effect turns out to be large or zero. If the authors want to ensure power for smaller effects, the BF criteria for stopping could be slightly increased, which would be more efficient than using a large minimum sample size.

Given the uncertainty around the ESs from the previous study, we chose to err on the conservative side and assume the data will be ambiguous, hence the higher numbers. This pessimistic expectation is also driven by the fact that we can only test a low number of trials per participant due to the low number of images in the database we are using (see our reply to Reviewer 2 below). We understand that this sampling approach may not be optimal in terms of number of participants tested. However, it allowed us to capitalize on the availability of students on campus for data collection and will hopefully reduce the analysis time. We now acknowledge in the article that this approach may not be the most efficient and was chosen in response of internal constraints:

"We appreciate that the lower boundary of the sample size is quite high and may not be as efficient as starting with a lower boundary in terms of participant time. This strategy was chosen due to internal timeline constraints requiring collecting more data upfront, but also justified by our simulations of effect sizes which suggest a high number of participants may be needed (see Effect Size section below)."

Minor points

1.9) I am not sure that abbreviated labels "C1", "C2" etc are needed. Descriptive labels could be used ("Fully Unconscious", "Mostly Unconscious", etc) without adding too much clutter in the text. Or could use "U" and "C" in the abbreviations to make it easy to remember which are unconscious vs conscious, "U","MU","MC","C", or "U1", "U2", "C2", "C1".

We have now changed all labels to U/MU/MC/C.

1.10) This topic sentence in the introduction is awkward: "Another relevant literature is the one referring to longer-term learning effects, and pertains to the increase in accuracy following repeated exposure to some stimuli over time." I suggest re-wording in a simpler manner, and maybe breaking off the second part to a new sentence.

The first part of the sentence has now been shortened and hopefully flows better with the second part as well:

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.

1.11) Another line that could be simplified: "In a conceptually similar context to that adopted by Chang and colleagues (2016), we aim to study whether the visual system can organise two-tone images into meaningful percepts after masked greyscale image exposure." Maybe something like this: "Using a similar method as Chang and colleagues (2016), we tested whether the visual system can organise two-tone images into meaningful percepts after masked greyscale image exposure."

The sentence has now been re-worked and simplified, thank you for the suggestion!

"Using a conceptually similar method as Chang and colleagues (2016), we test whether the visual system can organise two-tone images into meaningful percepts after masked greyscale image exposure."

1.12) The use of catch trials is listed as a different in method, but Chang et al (2016) also had catch trials. Are the catch trials different in the proposed study different?

There were indeed catch trials in Chang et al. (2016), however we argue that their catch condition was not sufficient to provide a suitable control for possible guessing. We explain this in the sections of the introduction on limitations to the original study and overview of proposed research, on pages 5 and 6:

"one cannot be certain that any genuine perceptual disambiguation occurred. Indeed, we 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 are 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."

Review by anonymous reviewer, 01 Feb 2023 16:03

1A. The scientific validity of the research question(s).

2.1) The authors present a sound argument for the validity of their research question. A previous publication (Chang et al., 2016) argued that grey-scale stimuli rendered unconscious by backward masking at an SOA of 67 ms could improve disambiguation of their Mooney (thresholded) counterparts. The grey-scale stimuli were categorised as 'unconscious' if the participant reported that they could not recognise the image. The present authors question whether the grey-scale stimuli were unconscious, as the previous authors did not control for response bias (where a participant may be unwilling to report that they could recognise an image though they consciously perceived it).

I would offer that, in addition to the arguments already made, the authors may wish to note that 67 ms is quite a long time for backward masking of visual stimuli. Bacon-Macé and colleagues (2005) show 85% correct performance in discriminating whether a natural scene contains an animal with an SOA of 44 ms and a much stronger mask (and accuracy was above chance at 12 ms SOAs). The weaker masks used by Chang and colleagues, and the proposed mask in this study, can be compared to RSVP studies, where performance above 75% correct can be achieved a stimulus presentation duration of 13 ms (even when the categorisation decision is only indicated after stimulus presentation; Potter et al., 2014). Based on this, it is very unlikely that the manipulation presented in Chang and colleagues

2016 experiment resulted in 'unconscious' stimuli, given this previous research showing participants can make quite accurate decisions about the contents of images presented at much shorter durations with stronger masking.

In the introduction, we chose to focus on the logical limitations of the Chang and colleagues' paper. We acknowledge that the concern raised about the SOAs is valid, and we plan to mention it, along with the suggested references, in the discussion.

1B. The logic, rationale, and plausibility of the proposed hypotheses, as applicable.

2.2) The authors' hypotheses are multifaceted. My interpretation is that the main aim is to test whether unconscious grey-scale stimuli can improve identification of their Mooney counterparts when using a combination of objective and subjective measures. The secondary hypotheses include comparing different criteria for consciousness, and the effect this has on whether Mooney identification can be considered significantly improved following exposure. Overall, these hypotheses are logical, rational, and plausible, and the results will be interesting whether the null is accepted or rejected.

It was often not clear what the authors meant by 'unconscious'. Most often, they do not qualify, and readers might presume they mean 'undetectable'. Frequently, they use the term 'conscious recognition', yet presumably they aim to present participants with images they have never seen before, and so even if the participant were fully conscious of the image, they would not recognise it on the first presentation. Sometimes they describe the phenomenon in terms of 'contents' (as they use in the description of the PAS ratings to observers). They should be careful here too, about the interpretation of what counts as 'content': would it be sufficient to have information about approximate figure/ground segmentation, or a general theme such as 'animal', or perhaps if the observer could tell whether the image was presented upright or inverted? It is also unclear if the authors presume an image could be detected (the observers are conscious of the presence of the image) while the 'content' is 'unconscious' – this might be important for some readers to understand whether the authors' criterion for consciousness matches their own.

It is true that we do not mention an all-encompassing clarification of what we mean by unconscious. This is because any theoretical stand we might take as to whether we think 'unconscious' means 'undetectable' or something else would have to be reflected in a specific measurement. As the reviewer points out, one of the aims of the paper is to evaluate precisely how/whether conclusions might change under different definitions of 'unconscious' as reflected in measurements, all which have been used previously in the literature (i.e., some readers find that our unconscious (U) condition aligns best with their definition, while others might find mainly unconscious (MU) preferable, while others might think that only using objective measures will lead to a satisfactory 'unconscious' condition).

We have added on page 2 a clarification about 'recognition' as this has been used by Chang and colleagues, but we have now replaced all the mentions of 'recognition' in our design with 'identification' to address the Reviewer's point. We have also added a clarification in the participants section about their eligibility: indeed, only participants who have never seen the images before are eligible.

We do not instruct participants specifically about what they should interpret as 'contents', nor at what level of specificity they should categorize the images in Experiment 1, but the instruction stage shows an example image and an example answer ('a deer'), which sets the expectation that their answers should be more specific than broad superordinate categories. Our text now mentions the above and

clarifies that, in our scoring of the answers in Experiment 1, 'animal' is considered an incorrect answer, since most of the images depict animals.

1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable).

2.3) Given the previous literature and the pilot data, it is not clear that the proposed backward masking paradigm is strong enough: participants report seeing a brief glimpse of the content of the image on about 50% of trials at the short SOA. The authors could increase the contrast of the mask to get stronger backward masking (Bachmann & Francis, 2014).

Since most of the data has already been collected (as specified in the PCI RR submission form), we are unable to change the contrast of the mask. The percentage of trials in the Short SOA rated with PAS1 (66% in Pilot 2, 60% in Pilot 3) might seem low, but we believe this is not an issue. Firstly, it is common occurrence that there is some degree of noise in how the participants use the PAS – for example, Jimenez et al. (2018) report that even in catch trials, containing only masks but no stimuli, participants still answered with PAS above 1 in around 25% of trials. Increasing the contrast of the mask in our study might likely result in some additional trials being rated with PAS1 instead of PAS2, but it will not bring the percentage close to 100. Secondly, in our Pilots 2 and 3, the overwhelming majority of trials not rated with 1 were rated with 2. These trials, as shown in Figure 7 (incorrectly labelled before as Figure 8), would form part of the mainly unconscious (MU) condition, so trying to reduce the percentage of trials in this category by employing stronger masking would undermine the analyses, and the broader aim of the paper to assess whether results change under different definitions of consciousness.

2.4) On that note, the example mask in Figure 4 looks as if it has broad vertical columns, it does not look like a phase scrambled version of the target stimulus as described.

Thank you for spotting this mistake. We have now updated the mask in the figure.

2.5) There are not enough trials to perform the analysis. The authors seek to test for an increase in disambiguation at C1 (PAS 1 + incorrect + short SOA) against catch trials. With 24 stimuli, half catch, half long SOA, half rated PAS 2 at the short SOA, that leaves 3 trials at C1 per participant. Based on Chang et al., 2016, they expect disambiguation to rise from ~2.5% to ~5%. Chang and colleagues had ~15 trials per participant, with 5% disambiguation this means that ~20/25 participants were able to disambiguate 1 trial each (20 trials out of the total 375). Even with 120 participants (the maximum), the authors will have a total of 360 trials, 18 correct, meaning 18/120 participants get 1 trial correct each. The number of trials should be at least doubled. The authors could fit double the number of trials in a similar amount of time by reducing the mask duration (500 ms should be more than sufficient) and the fixation duration (or some of the beginning of the trial – Figure 4 suggests there is 0.5s blank, 1s fixation, 0.1s blank, 0.2s fixation – 0.2s blank followed by 0.5s fixation should suffice). Increasing the masking effect (by increasing the contrast of the mask) should also help to get more trials at PAS 1 at the short SOA.

Although we fully agree with the clear benefit of having more trials, we cannot increase the number of trials unfortunately. The stimulus set produced by Teufel and colleagues (2015) underwent significant piloting and refining to arrive at pairs of two-tones and templates that were not too easy to disambiguate right away before template exposure, and not too difficult to disambiguate after template exposure (under reasonably long exposure times). We aim to compensate this low trial number with up to almost 5 times the sample size in the original study. Our Bayesian analyses would be able to inform us if there is not enough data (through an inconclusive Bayes Factor) even with 120

participants, although as described in response to Reviewer #1 above, we believe the study is sufficiently powered for detecting a medium effect size of 0.5.

2.6) The pilot data suggests that participants are remarkably good at the task in the pre-exposure phase, with accuracy ~15% correct free naming identification (Figure 6). This is much better than reported in Chang et al., 2016. It could also be problematic that there is quite a substantial increase in catch trial performance, considering the small number of trials. All this will make it even more difficult to get good measures of performance with only 3 trials per participant. The authors could get a better estimate of the likely statistics by dividing pilot participants' data as if they were different participants (the pilot has 19 participants with 12 trials = 228 trials total, can be divided into 76 participants with 3 trials each, to estimate whether the effects could be reliably detected with so few trials).

The higher rate of free naming identification in our study compared to Chang et al. (2016) could be due to differences in how conservative the accepted answers list was – which we cannot evaluate since their list is not available. In any case, we do not view that as a concern because, as shown in Figure 6, the rate pre-exposure was very similar between catch and experimental trials, and ultimately our tests rely on the differences between these two conditions. A substantial increase in catch trial performance is again not a concern for us, because it could be explained by a number of factors that could have contributed to the increase (for example, as described in the paper, category-relevant visual or semantic information in the catch greyscales, or spontaneous disambiguation). Critically, in Pilot 1, the exposure time was much longer than the 17ms in the main experiment, and it was not masked – it is therefore expected that Pre-Post changes would be much higher.

2.7) It is worrying that so many of the long SOA trials were rated as PAS 2 in the pilot data. This could indicate that participants do have some bias, or that they are relying on some other cues to make their ratings. The example grey-scale images look as though they have different RMS contrast and spatial frequency properties. The catch-trial peacock appears much more difficult to see (in the pdf, long duration) than the 'main' trial peacock. I wonder if some participants might be using these cues to separate out their PAS ratings over clearly visible stimuli. The authors could check this in their pilot data, they should also report on the variability of low-level stimulus properties, or even control the low-level stimulus properties.

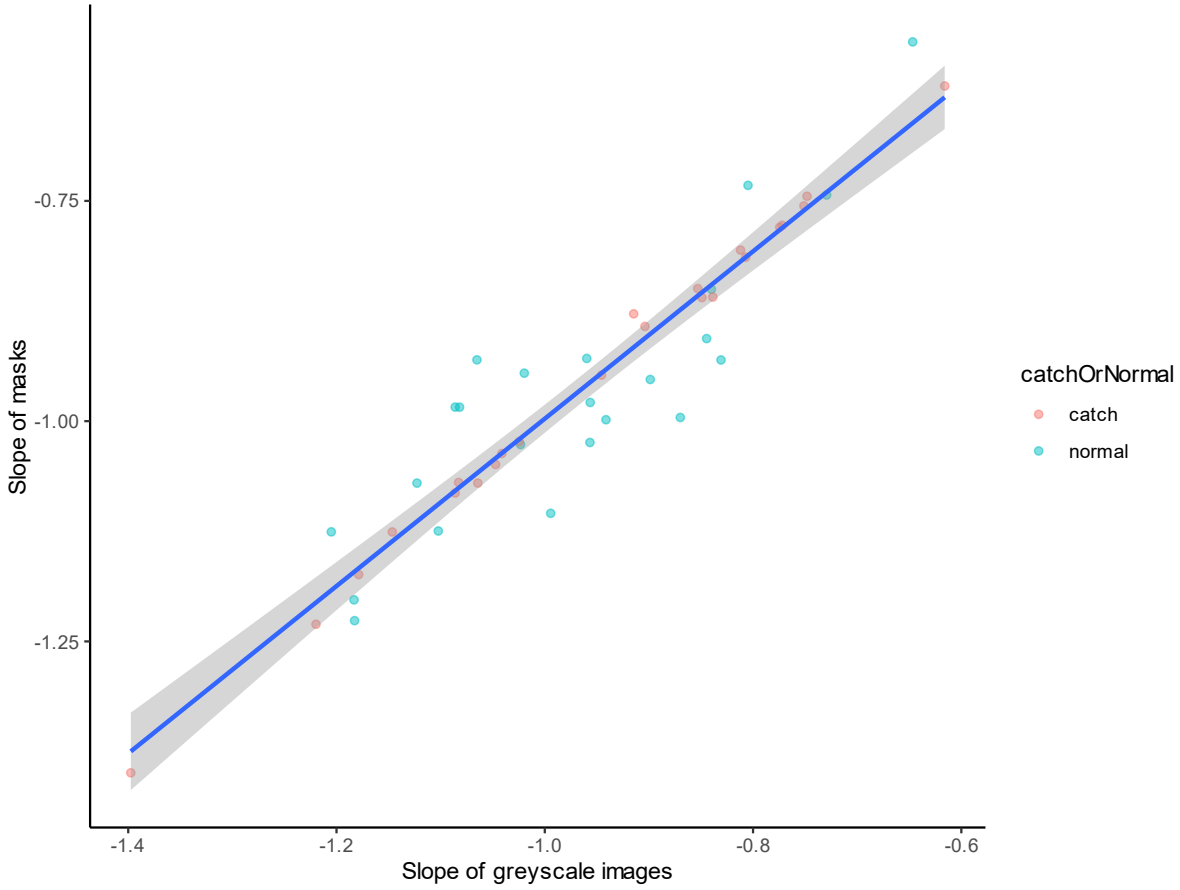
While it is true that a high number of trials were rated as PAS2 ("a brief glimpse"), we believe this is to be expected given that the task is quite difficult – the display time is 17ms and the images are complex natural scenes, which does not really allow for much clarity. In any case, in the Long SOA conditions at least half of trials in Pilots 2 and 3 were correctly identified, suggesting that even if participants said they only briefly saw the content, that was often sufficient to result in a correct answer.

Thank you for the suggestion to consider the low-level properties of the images. We conducted the analyses described below and concluded that there are no systematic differences in our image set that could impact the results, as detailed below.

We computed the RMS contrast, luminance (using the SHINE toolbox), and edge density ("Sobel" parameter in MATLAB) for each image after resizing it to the dimension that participants would see (121x121 px), and compared them between the Catch and Test sample, for each image pair (Bayes paired t-tests, all error values under 0.1%). We found BFnull = 5.65 for RMS, BFnull = 5.07 for edge density, and BFnull = 6.21 for luminance – therefore evidence for no systematic differences in these properties. For spatial frequency, we calculated the radial average of the Fourier spectrum for each Catch and Test two-tone and mask image, log-log plotted the amplitude spectra, and extracted the

slope value (e.g., Tolhurst et al., 1992; van der Schaaf & van Hateren, 1996). We used a lower limit of 4 cycles/image and an upper limit of 31 cycles/image (image size/4), in order to avoid our estimates being biased by unreliable low/high frequency values.

Hansen and Loschky (2013) argued for an "amplitude slope similarity principle" in a very similar design to ours (gist perception paradigm with 12ms target natural images and phase-scrambled masks), suggesting that the effectiveness of the masking is influenced by the degree to which the amplitude spectra of the target and masks differ, rather than their absolute values. To test this, we entered the slope values into a mixed model with Trial Type (Catch or Main) and Image Type (masks or templates) as fixed effects and image pair as random intercepts. We found BFnull = 3.47 (error = 1.2%) for an interaction between Image and Trial Type, as well as evidence against any main effects (Image Type BFnull = 4.7, error = 0.96%; Trial Type BFnull = 3.65, error = 1.23%). Indeed, all models showed evidence for the null, with the strongest model containing both predictors and the interaction, BFnull = 58.41, error = 2.14%. All tests used a wide prior scale. Slope values for each image and its corresponding mask is included below.



2.8) The effect size calculation compares accuracy to 0, and so does not match the main hypothesis – compare 'unconscious' exposure to catch trials (~5% to ~2.5% correct). However, this estimation is not used for anything, so it could be removed.

We have now removed the table, and justified further our choice of ES and why deriving an ES from the previous paper is difficult and possibly not as informative as hoped – please refer to answers 1.7 and 1.8 in response to points raised by Reviewer #1.

1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses.

2.9) The methodology is highly detailed, however there are some typographical errors and ambiguities. Experiment 1 stimuli description lists 23 images + 9 attention checks, the table in Appendix 2 lists 24 images, the detail about attention checks lists 9 visible + 6 absent, the detail in point 2 of participant rejection suggests 6 visible.

There are indeed 23 experimental images + 9 attention checks, as mentioned in Experiment 1 stimuli description. A 24th image was added only to achieve equal block lengths - Appendix 2 shows the block structure factoring in equal block lengths. This 24th image will be removed from all analyses. We mention this in the stimuli description, however we have now also added a note in Appendix 2 clarifying this.

Thank you for noticing the inconsistency in reporting the attention checks. There are 9 visible attention checks, this has been modified and the exclusion criteria changed accordingly (1/9 instead of 1/6).

2.10) The authors may wish to specify that their hypothesis (and analysis) is one-sided (they expect increased performance after exposure, but presumably a decrease in performance would be evidence for the null).

We now follow the reviewer's recommendation for one-sided tests and updated the paper accordingly (page 13).

1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).

Yes, the catch trials should be sufficient control.

Minor points:

2.11) There are some typographical errors, e.g. page 5 1st line "Chang et colleagues" and last line "undistinguishable". "for which likelihood of" to "which the likelihood" (abstract)…

Thank you for flagging these, they have now been corrected.

2.12) Figure 6 does not have panel labels.

Panel labels have now been added.

2.13) Perhaps there is a better name for 'normal trials'/'main trials' – for example 'relevant exposure trials' or 'test trials'.

The 'main' trials have now been re-named as 'test' trials, and all mentions of 'normal' trials (equivalent to main/test) have now been removed for consistency.