We would like to thank the reviewers for their time and effort. Their insightful comments and suggestions have not only improved our manuscript, but also enriched our understanding of the subject, and for this, we are very thankful. Below are our point-by-point responses (in blue font) to the reviewers' comments (in black font). Please note that the page numbers refer to the revised manuscript with tracked changes.

Anonymous reviewer 3

This version has already been reviewed and the current RR looks fine to me. My main concern is that the alternative hypothesis seems rather like a strawman. That is, if the results aren't obtained, the authors claim that "The assumption that masked priming paradigms with and without trial-by-trial judgments of prime visibility lead to identical priming effects could be shown wrong. " But who hold this assumption? Why would anyone do so? Even if it were in fact a 'strawman' of sort, I suggest that the authors should still better motivate it, and show/argue more explicitly that this has at least been an implicit assumption made in some previous research. Obviously, we all make implicit assumptions, out of convenience or sheer laziness. So, even if in reality I think very few people actually strongly hold this assumption in earnest, probably a case could still be made. How interesting the resultant paper is seems to hinge on how strongly the authors can make this case, to show why these results would really matter, rather than just trivially expected by everybody.

Thank you for bringing this to our attention. We originally put the claim that "The assumption that masked priming paradigms with and without trial-by-trial judgments of prime visibility lead to identical priming effects could be shown wrong" into the design template for the sake of completeness. The claim does not present the motivation for our study. We have therefore revised that section of the design template, ensuring the claims are more clearly differentiated:
"The assumption that task 2 in general, and more specifically, its characteristics, do not affect task 1 within a dual-tasking paradigm could be shown wrong. The assumption that concordant input/output modality pairings lead to less interference than not concordant input/output modality pairings could be shown wrong." (p. 34)
Also, we would like to argue that our motivation is actually to show that masked priming paradigms with and without trial-by-trial judgements of prime visibility do differ in their priming effects, because task 2 (the prime visibility assessment) influences task 1 (the priming task, the indirect task). The alternative hypothesis is therefore that they don't.

Markus Kiefer

The authors present a stage 1 PCI Registered Report of an envisioned study, in which they systematically investigate the demands associated with the dual-task structure of a masked priming paradigm with ive visibility judgments. Response modality and task complexity is systematically varied across experimental blocks, while behavioural data and event-related potentials are collected.

The topic of the study, the dual-task structure of a masked priming paradigm with subjective visibility judgments, is interesting and timely. Overall, the study is well designed and described with sufficient scrutiny. However, several critical issues as outlined in detail below should be addressed.

1.) First of all, the authors should clearly indicate from the beginning that their study is specifically focused on responses priming (lines 2-7). In the following paragraph, they might want to describe in more detail the difference between semantic priming and response priming (e.g., Martens, U., Ansorge, U., & Kiefer , M. (2011). Controlling the unconscious: Attentional task sets modulate subliminal semantic and visuo-motor processes differentially. Psychological Science, 22(2), 282–291.)

We added the following paragraph: "We will be exploring the concept response priming, utilising arrows as primes and targets, for which priming is the result of visuomotor processes. In semantic priming, in contrast, priming stems from access to word meaning (see Martens et al., 2011 for more detail)." (p. 3)

2.) The dual task situation and its impact on priming-related processes have been intensively discussed in Kiefer, M., Harpaintner, M., Rohr, M., & Wentura, D. (2023). Assessing subjective prime awareness on a trial-by-trial basis interferes with masked semantic priming effects. Journal of Experimental Psychology: Learning, Memory, and Cognition, 49(2), 269-283. https://doi.org/10.1037/xlm0001228.). In particular, these authors have proposed five mechanisms via which visibility ratings could alter priming-related processes, some of them are particular relevant for response priming: attentional focus on perceptual prime features, reduction of attentional capacity and response-related interference. Most interestingly, while attentional focus on perceptual prime features would enhance response priming, the latter two mechanisms would reduce response priming. Depending on the net contribution of these mechanisms, either enhanced priming or reduced priming during trial-wise visibility ratings is observed, possibly interacting with task complexity. I recommend to describe these proposed mechanisms and to include them in their predictions. These suggested mechanisms are important because according to current research trial-wise visibility ratings seem to enhance the magnitude of response priming, while they reduce the magnitude of semantic priming.

Thank you for the suggestion. We added the following paragraph to our introduction: "Interestingly, current research suggests that trial-wise prime visibility ratings lead to a decrease in semantic priming, as observed in the studies mentioned above, but to an increase in response priming (e.g. Biafora & Schmidt, 2022). Kiefer and colleagues (Kiefer, Harpaintner, et al., 2023) describe mechanisms altering prime-related processes that offer an explanation. The trial-by-trial awareness rating may lead to to (1) an emphasis of an attentional focus to perceptual features of the prime, to (2) a reduction of attentional capacity or a plus of attentional demands as compared to a single-task situation, and to (3) response-related interference due to an increase of non-decisional process components like response-related processes. While the first mechanism would enhance response priming, the latter two would reduce it, and therefore, depending on the net contribution of these mechanisms, trial-wise visibility ratings can either lead to enhanced or reduced response priming as compared to a single-task situation (Kiefer, Harpaintner, et al., 2023)." (p. 5)
We also added these mechanisms to formulate predictions about our manipulations' influence on the observed response priming effects and changed our hypotheses accordingly. (p. 22)

3.) The authors should improve their description of the relation between subjective and objective measures of awareness (lines 193-197). Firstly, they missed to describe recent empirical work, which demonstrates a convincing convergence of subjective and objective measures indicating that both measures can validly capture the content of awareness (Kiefer, M., Fruehauf, V., & Kammer, T. (2023). Subjective and objective measures of visual awareness converge. Plos One, 18(10). https://doi.org/10.1371/journal.pone.0292438).
Secondly, they missed to refer to a recent review paper highlighting the conditions, under which a convergence of divergence of objective and subjective measures is found (Kiefer & Kammer, 2024). This paper also questions the claim that "subjective ratings are argued to be better suited to accurately grasp the content of phenomenal consciousness as compared to the standard objective measure" (lines 195-197)
Thirdly, the reference to Kiefer et al. (2023) within the context of the statement "subjective ratings are argued to be better suited to accurately grasp the content of phenomenal consciousness as compared to the standard objective measure" is wrong. If anything, both Kiefer, M., Harpaintner, M., Rohr, M., & Wentura, D. (2023) and Kiefer, M., Fruehauf, V., & Kammer, T. (2023) demonstrate a convergence of measures. The authors should instead refer to Overgaard, M., Rote, J., Mouridsen, K., & Ramsoy, 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/DOI 10.1016/j.concog.2006.04.002, Sergent, C., & Dehaene, S. (2004). Is consciousness a gradual

phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. Psychological Science, 5(11), 720-728.

Thank you for bringing this to our attention. First and second point: we decided to dedicate a new section to our reasoning behind our choice of the PAS, which includes the convergence of subjective and objective measures (see paragraph "Choice of visibility measure", p. 13).
Third point: we corrected the argument made here and added a suitable reference. (p. 12)

4.) Lines 198-200: Kiefer and colleagues did not compare difference subjective measures. The appropriate reference is: 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.

Thank you for catching this. We fixed the citation in the manuscript.

5.) Power analysis (lines 276-279) and statistical analysis (438-443): I do not understand why the authors want to use one-tailed paired t-tests on priming scores and not a 2x 2 x 2 rp measures ANOVA with the factors congruency, response type and complexity. Comparison of priming scores does not reveal whether priming is reliable at all (e.g. larger than zero). Multiple t-tests also inflate false discovery rate, if not controlled for, and do not allow to test interaction effects. Multiple t-tests controlled for false discovery rate could be used as a post-hoc analysis, when interactions are significant. I suggest to move the ANOVA from the exploratory analysis to the main analysis.

Thank you for pointing this out. We will use ANOVAs instead of multiple t-tests and we changed the manuscript accordingly:
"We will conduct one repeated-measures three-way ANOVA comprising the factors response modality (vocal vs. manual), response complexity (high vs. low), and prime-target congruency (congruent vs. incongruent) to test for RT differences, as well as a repeated-measures two-way ANOVA comprising the factor task type and congruency to test for RT differences between single and dual-task. Multiple t-tests controlled for false discovery rate will be used as post-hoc analysis." (p. 24)
We changed the power analysis accordingly: "In G*Power 3.1.9.7 (Faul, Erdfelder, Lang, & Buchner, 2007) we calculated the sample size for the repeated measures 2x2x2 ANOVA using a medium effect size f (0.25; partial eta squared = 0.06) for the main effects (Cohen, 1988). Assuming a mean correlation between repetitions of 0.5, we determined that for f = 0.25, and alpha = 0.05, a sample size of N = 34 was required to achieve a power of 0.80 (measurements: 2; groups: 1)." (p. 16)

6.) Some points with regard to the methods should be clarified:

a) Are the authors sure that the prime-mask/target SOA of 8 frames (line 366) renders the prime invisible in all participants?

Thank you for pointing out this issue. We initially chose 8 frames, as longer SOAs lead to larger priming effects. Since we are interested in investigating influences of a dual-task in a paradigm of low visibility rather than complete unconsciousness, we were not primarily worried with rendering the prime invisible for all participants. Based on the reviewer's comment, we re-analysed the data from a pilot experiment where we used 2, 5, and 8 frames. Indeed, the SOA of 8 frames was associated with rather high PAS levels (mean PAS: 1.66; 0-3 PAS). We therefore decided to use a 5 frame SOA instead (appr. 59 ms) as it more adequately complies with our requirements of low visibility (mean PAS: 1.06; 0-3 PAS).

b) As the electrode gel is simply injected in active electrodes, the gel typically has no abrasive (line 336) properties on the scalp.

We have removed the reference to the abrasive gel.

Thank you for drawing our attention to this. We plan on following a recommendation made by Farrens and colleagues (2021) and changed the manuscript accordingly to:
"For light skin abrasion which helps reduce electrode impedances, participants will be asked to comb their hair with a plastic comb, concentrating on the scalp (Farrens et al., 2021)." (p. 20)

Farrens, Simmons, & Luck, S. J. (2021, November 16). *Electroencephalogram (EEG) Recording Protocol for Cognitive and Affective Human Neuroscience Research*. https://doi.org/10.21203/rs.2.18328/v4

c) lines 345-350: As the authors record EEG with 32 electrodes, the dimensionality reduction to 64 dimension is unclear. The authors should indicate the nature and number of initial dimensions. It is also not clear whether the PCA and ICA is calculated to remove ocular artifact components. If yes, this should be explicitly mentioned right from the beginning. If not, the purpose of these transformation should be explained. It is also not clear why the PCA and ICA is not calculated on continuous EEG data to better capture ocular artefacts.

Thank you for drawing our attention to this mistake. We corrected it as follows:
"Independent-component analysis (ICA) will be performed on the concatenated single-trial EEG data, using the extended INFOMAX algorithm as implemented in EEGLAB (Bell & Sejnowski, 1995). The resulting 32 ICs will be automatically classified using the ADJUST toolbox (Mognon et al., 2011) and rejected if classified as artifact (i.e., eye blink, eye movement, and generic discontinuity)." (p. 19)
Please note that Marco Buiatti, one of the authors of the ADJUST toolbox, confirmed to us that this ICA-based toolbox can be run on both the continuous EEG data and the concatenated single-trial EEG data.

d) The authors should indicate the type of monitor, its refresh rate and timing accuracy. Timing accuracy should be explicitly measured and controlled for, because primes are only presented for two frames (line 365).

We described the type of monitor we use in the Section Apparatus and Stimuli: "The participants will be seated in a dimly lit room in front of a Samsung Samtron 98PDF CRT-Monitor (1280 x 1024 pixels, refresh rate 85 Hz, grey: 31 cd/m2) at a viewing distance of approximately 60 cm." (p. 17)

c) Why is EEG sampled with 1 kHz? 500 Hz might be sufficient for the authors' purposes.

We agree that 500 Hz will be sufficient and we changed the manuscript accordingly. (p. 19)

7.) The statement in the abstract is wrong: "In masked priming, the prime's visibility is typically assessed with a subjective measure on a trial-by-trial basis". This is not true. As described in Kiefer et al. (2023), in masked priming experiments prime visibility is typically assessed in a separate session after the priming phase, in order to avoid interference of the visibility judgments with the priming effect.

Thank you for bringing our attention to this. We changed this sentence from the abstract to: "In masked priming, the prime's visibility can be assessed with a subjective measure on a trial-by-trial basis." (p. 2)

8.) Line 67: The reference of Kiefer et al. 2023 within the context of response priming and arrows as stimuli is wrong, because these authors investigated semantic priming.

We corrected the citation in the manuscript.

9.) The references are not always complete, for instance in line 620:

Mattler, U. (2003). Priming of mental operations by masked stimuli. 167–187.

Thank you for catching this one. We completed the reference in the manuscript.

10) When referring to Kiefer et al. (2023), please add "a" or "b" to distinguish between two articles published in 2023 by this first author.

Thank you for pointing this out; of course, both references need to be distinguished within the manuscript. We did not use "a" and "b" but differentiated between both sources by putting "Kiefer et al., 2023" for one and "Kiefer, Harpaintner et al., 2023" for the other (APA 7th).

Thomas Schmidt

Review of Registered Report "Probing the dual-task structure of a metacontrast-masked priming paradigm with subjective visibility judgments", by Charlott Wendt and Guido Hesselmann

Reviewer: Thomas Schmidt

The authors are investigating a direct measure of priming and an indirect measure of awareness in a masked priming paradigm. Here they are planning a study investigating the consequences of administering direct and indirect tasks simultaneously, i.e., as a dual task, compared to sequentially, i.e., as single tasks in separate blocks. The stimulus sequence consists of a 24-ms arrow prime, a single 94-ms SOA, and a 106-ms arrow target (congruent or incongruent with the prime). Fixation onset is variable. The indirect task is speeded discrimination of target direction (to measure priming), and the indirect task is a rating on a custom-made visibility scale (modified PAS). Different blocks (all performed in a single session) vary the complexity of the rating (2 or 4 categories) and the response modality of the rating (voice or keypress). Two additional blocks measure target discrimination in a single task as well as a forced-choice objective measure of prime discrimination.

By and large, this is a sound research plan and I am looking forward to seeing the results. As a reviewer of a registered report, I see my role as suggesting improvements in methodology and in the analysis plan while resisting the urge of imposing my own idiosyncratic preferences on the researchers.

MAJOR POINTS

- For many reasons, I would wish for a manipulation of the SOA, but I see that this would require multiple EEG sessions.

We would like that, too. But we made the decision here to save on the time participants have to spend in the laboratory.

- In my opinion, the question is not only whether the dual tasks interferes with the direct task, but also what it does to the indirect task (priming). One of the findings in the Biafora & Schmidt paper was the loss of time-locking between RT and prime onset under multitask conditions. As response time increases under task load, more variance is introduced and the time-locking may suffer, which is an indication that the bottom-up, feedforward link between stimulus and response is no longer effective. Even with the fixed SOA, the authors could use the variable fixation/intertrial interval to take a look at time-locking to the stimulus (not only for RT, but also for EEG). Other aspects of the RT distributions would be relevant as well: do the distributions become wider under dual tasks, are priming effects still observed in the quickest responses, and are there fast errors (i.e., are errors as fast as the fastest correct responses)?

It is our main concern to look at what the dual-task does to the indirect task (priming), which is why we are looking at RTs and priming effects in task 1 (speeded two-choice identification task).
Using the variable fixation/inter-trial interval to look at time-locking of the stimulus is an interesting point and we might add this to our exploratory analysis, but would at this point rather not preregister it. We do agree that other aspects of the RT distributions should be relevant and will therefore add these to our exploratory analysis: "Regarding the RTs, we will also investigate other aspects of the distribution, for we expect distributions to be wider for dual as compared to single-task, and are interested in whether priming effect are still observed in the quickest responses, and whether there are fast errors, i.e., whether errors are as fast as the fastest correct responses." (p. 24)

- I was surprised that Block F introduces a new (objective) prime discrimination task, but none of the two rating scales. While discrimination performance is certainly interesting, wouldn't it be relevant to see whether the dual task changes the visibility ratings in any way? One concern that we had in our paper was that participants may monitor response conflict in the first response and use that to infer the identity of the prime. For instance, response errors can mostly be attributed to incongruent primes, so prime identity may be guessed from response accuracy in any trial. We found no evidence that participants used that strategy in our experiments, but of course this could be different here, especially with arrow stimuli.

This is a rather interesting point and we are grateful for you drawing our attention to it. Our focus is, however, on the influence on dual-task on the priming task. We might therefore add this to our exploratory analysis, but would at this point rather not preregister it, for we feel it to be beyond the scope of the current manuscript.

MINOR POINTS

- Have the stimulus specifications been tried out yet? It could be that with foveal presentation, the relatively strong prime is difficult to mask. Even if masking works well, the authors have to anticipate that participants will differ markedly in their masking effects.

Please see our response to M. Kiefer's comment regarding this issue above.

- Just as an aside: it is not completely trivial to explain why priming effects should increase with RT; there are certainly side conditions for that. Thinking along the lines of an accumulator model, reducing the input to the counters would lead to slower accumulation, and the accumulation functions would hit the RT boundary later and under a flatter angle. That would predict both longer RT, more variance, and larger priming effects. On the other hand, the preactivation by the prime would also be weaker, and that would decrease the priming effect. It all hinges on the relative strengths of prime and target.

This is a very valid point; thank you for drawing our attention to this. We updated our prediction for the priming effect by taking the three mechanisms into account by which trial-wise prime visibility ratings could alter prime-related processes as discussed by Kiefer, Harpaintner et al., 2023. (p. 22)
See our response to reviewer M. Kiefer.

- p. 10: From our theoretical perspective (Schmidt & Biafora, 2024; Schmidt & Vorberg, 2006), there is no sense in saying that one measure is "more exhaustive" or "more exclusive" than another, because exhaustiveness and exclusiveness are all-or-none properties and often unattainable for realistic measures. It is much more sound to discuss the similarity of their criterion contents and whether these contain the critical feature. In this experiment, the critical feature is the prime direction, and it is certainly contained in the criterion contents of either task (as already reflected in the authors' wording of the rating categories).

We removed this statement and put our reasoning behind our choice of measure into a new section, see “Choice of visibility measure”. (p. 13)

- Shouldn't the predictions for the EEG results include LRPs? Those are the primary means to look at response priming effects beginning with Eimer & Schlaghecken and Leuthold & Kopp. If larger RTs lead to more priming, what happens in the LRP? It should become more stretched out in time and lose its time-locking, shouldn't it?

Thank you for the suggestions. This, too, is an interesting proposal and we might add this to our exploratory analysis, but would at this point rather not preregister it.

- Just as a remark to the previous review: Predictions for EEG effects do not necessarily require time windows from pilot data. There is also the strategy of defining landmarks in the waveform (e.g., onset of an LRP, time and amplitude of a peak) and use jackknifing to perform the statistical test on those landmarks (which is really easy, see Miller & Ulrich, 2001). Because the overall shape of the LRP is relatively clear beforehand, most researcher degrees of freedom are eliminated this way.

Thank you for pointing this out. We decided to use recommendations made by Kappenman et al., 2021 concerning the time windows (300 to 600 ms for P3) and changed the manuscript accordingly (p. 25). As suggested, we will use the jackknife approach.

Kappenman, E. S., Farrens, J. L., Zhang, W., Stewart, A. X., & Luck, S. J. (2021). ERP CORE: An open resource for human event-related potential research. *NeuroImage*, *225*, 117465. https://doi.org/10.1016/j.neuroimage.2020.117465

- The power analysis combined with 60 trials per cell and subject is convincing to me and ensures the measurement fidelity of the experiment even if based on t-tests and not the actual RM-ANOVA. Here's my reasoning. If RT distributions have an SD around 60 ms and are based on 60 observations per condition, that implies that individual persons' standard errors around single datapoints are around 60/sqrt(60) = 8 ms, which is fine measurement precision. It means that differences around 16 ms can be statistically resolved within an individual observer, which for me is a relevant psychophysical standard of data quality. And that's the point of a registered report, isn't it: to ensure the validity of the design and the quality of the measurement and then live with whatever it is the participants produce. If they are homogenic in their effects, a la bonne heure; if they are inhomogenic (but well measured), we have to report it all the more and try to find the explanation in subsequent research. In contrast, a formal power analysis of a multifactorial repeated-measures design is usually neither straightforward nor convincing. It can only be done by simulation (not G*Power), and the results are usually questionable because all assumptions about the critical effect x participant interactions are usually guesswork.

Thank you for this comment. Please note, however, that our revision now contains a power analysis for the planned 2x2x2 ANOVA.

Anonymous reviewer 2

This registered report proposes to study the effect of awareness measures on the obtained effect, focusing on behavioural data, and also adding electrophysiology as an exploratory analysis. I found the research question important and interesting, and I think the results will be impactful for future studies. I did have some comments/suggestions below, but I am certain that all of them can be addressed, such that the report could be accepted for publication.

1. When discussing objective and subjective measures, the authors mention Pereman & Lamy's results. This is great, but this description is missing other findings, suggesting that there is a difference between objective and subjective visibility (e.g., Stein et al., 2023, Plos Biology).

We added a section discussing the difference between objective and subjective measures. See paragraph "Choice of visibility measure". (p. 13)

2. I am not sure that the low-complexity PAS should still be called a PAS… It basically amounts to "see" vs. "didn't see", and the entire idea behind the PAS, at least the way I understand it, was to add the intermediate levels to allow a more refined and nuanced means of reporting. I would accordingly suggest changing the terms and say that the complexity of the subjective measure was manipulated, with high (PAS) and low (dichotomous) complexity subjective measures.

This is a very valid point and we therefore renamed our two-item measure to "dichotomous subjective measure" throughout the manuscript. E.g. "For the low-complexity condition, there will only be two items (…), and we coined this the dichotomous subjective measure." (p. 13)

3. P. 12, first sentence of the first paragraph (starting with "the latencies of earlier…") => I believe this sentence is not complete, unless I am missing something. Also, in the last sentence of this paragraph, it says: "whether the target-related P3b responses would show a differential and amplitude depending…" => I believe the "and" should be removed?

Thank you for catching this. The sentence needs to end like this: "The latencies of earlier sensory ERP components, such as the P1 and N1, have been consistently reported to remain stimulus-locked to both targets and show no postponement related to dual-tasking."
We removed the "and".

4. If I understand correctly, the power analysis was conducted based on a behavioral effect, yet this is also an electrophysiological study. Shouldn't the power analysis be conducted also on one of the P3b studies that found an effect (such that the bigger sample size would be chosen)? I appreciate that the EEG part of the study is exploratory, but it would be a waste of resources to collect all the EEG data and find no effect since the sample is not powerful enough. I saw the reply to a similar point in the first round, yet still think that this should be taken into account.

We agree with the reviewer that it would be a waste of resources if our EEG data analysis were clearly underpowered. However, we are quite confident that this is not the case. Assuming a medium-sized effect on P3b amplitude (e.g., a main effect of response modality), our sample size of N=34 would give us a power of 0.8 (see our revised apriori power analysis). Unfortunately, it is not trivial to rely on previous studies here, since - to the best of our knowledge - no study so far investigated the effects of response modality and response complexity on P3b amplitude in a dual-tasking situation.

5. The authors write: "The PAS will serve as the direct measure of prime processing"; I believe this is not fully accurate. I think it serves as a direct measure of prime visibility – the latter can be completely absent yet the prime will still be processed (this is exactly what the authors are aiming for), so it's a measure of visibility, not of processing.

We agree, and we changed the manuscript accordingly. (p. 18)

6. I didn't understand the sentence in the method (p. 15) saying that the authors will "use only a single SOA due to time constraints". Why even mention it as a variable if you only have 1 SOA?

Thank you for catching this one! That is indeed unnecessary. We changed the manuscript as follows: "Our experiment will hold a total of 10 conditions: congruency (congruent vs. incongruent) and block (A, B, C, D, E)." (p. 18)

7. 0.5 is a pretty high value for a high-pass filter. Why was that chosen?

We changed it to 0.1, as recommended by Luck, 2014. (p. 19)

Luck, S. J. (2014). *An introduction to the event-related potential technique* (Second edition). The MIT Press.

8. Given that the stimuli are presented in Figure 2, I found Figure 1 redundant.

That is a good point. We removed figure 1 from the manuscript.

9. Are the researchers planning to exclude trials in which visibility is higher than 0 where PAS is measured? If so, do they plan to account for RttM in any way? And how can they exclude the option that in the single task condition, where such trial exclusion does not take place, the predicted stronger effect does not stem from the inclusion of conscious trials? If they are not planning to exclude trials, how can they make sure that the effect do not reflect some residual conscious processing?

We do not plan to exclude trials in which visibility is higher than PAS>0. We cannot be sure that the effect observed in the single-task does not reflect some residual conscious processing. However, since the main purpose of our study is to investigate the influence of a dual-tasking structure itself, we are not too worried about rendering the prime invisible for every participant. Rather, we are interested in a paradigm of low or reduced visibility. We make this point in the section *Design* in our manuscript. (p. 18)

10. Why test the behavioural predictions with several t-tests rather than an ANOVA, followed by planned comparisons/post-hoc analyses? I saw the additional ANOVA at the end, but this feels redundant. I would simply start with it (unless there is a good reason not to). I saw a referral to this in the first round, but didn't quite understand the rationale.

We agree with this and replaced the t-tests with ANOVAs. Please see our reply to reviewer Markus Kiefer.

11. I didn't see any referral to a correction for multiple comparisons, although several ones are over the manuscript (see the following paper explaining why this is crucial: Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., & Golani, I. (2001). Controlling the false discovery rate in behaviour genetics research. Behavioural brain research, 125(1-2), 279-284). There are different methods one could use; I suggest adopting the latest tree-like method by Benjamini, which avoids an overly strict correction by taking into account the nested structure of these comparisons (Bogomolov, M., Peterson, C.B., Benjamini, Y., and Sabatti, C. (2021) Hypotheses on a tree: new error rates and testing strategies. Biometrika. 108(3), 575-590).

Thank you for pointing this out. We will correct for multiple comparisons using false discovery rate and changed the manuscript accordingly: “We will conduct one repeated-measures three-way ANOVA comprising the factors “response modality2 (vocal vs. manual), “response complexity” (high vs. low), and “prime-target congruency” (congruent vs. incongruent) to test for RT differences between conditions , as well as a repeated-measures two-way ANOVA comprising the factor “task type” and “congruency” to test for RT differences between single and dual-task. Multiple t-tests controlled for false discovery rate (FDR, (Benjamini & Hochberg, 1995) will be used as post-hoc analysis.” (p. 24) Please note that we will use the same approach for our ERP analysis.

12. EEG: wouldn’t it be better to average Fz Cz and Pz? Given that there’s no expectation for a difference between electrodes.

Thank you for the suggestion. However, we will look at all three electrode sites separately because we feel this is how it is usually done in research looking at dual-tasking paradigms.
“We will be using the outputs from the three midline channels Fz, Cz and Pz to isolate the P3b, as these are typically used in dual-tasking paradigms probing P3b (Aliakbaryhosseinabadi et al., 2017; Isreal et al., 1980; Kappenman et al., 2021; Kasper et al., 2014; Knott et al., 2003), and the average of both mastoids as reference (Kiesel et al., 2008).” (p. 25)

Kiesel, A., Miller, J., Jolicœur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *Psychophysiology*, *45*(2), 250–274. https://doi.org/10.1111/j.1469-8986.2007.00618.x

13. Selecting the time windows by means of inspection raises the concern of double dipping. Is there an independent way to define the time windows (e.g., using a subset of the data? Or a small pilot experiment?). Again, the reply to this point in the first stage review is not satisfactory, I’m afraid. Even if this is an exploratory analysis, it should be done properly, and I am not sure visual inspection is the best strategy here, I’m afraid.

Thank you for pointing this out. We did not choose time windows used in other studies before because we felt our design to be somewhat too novel for that. But we now decided to use rather general recommendations from Kappenman et al, 2021 for the P3:
“Statistical analyses will be calculated over mean amplitude and onset latency values in a time window recommended for the P3 by Kappenman and colleagues (2021): 300 to 600 ms. We will be using the jackknifed averages for mean amplitudes and, following a recommendation by Kiesel et al., (2008), we will combine jackknifing with the relative criterion technique with parameter 50% for calculating onset latencies.” (p. 25)

Anonymous reviewer 1
In this pre-register study, the authors aim to investigate the dual-task architecture in the study of unconscious processing using a metacontrast masking experiment and event-related potentials (ERPs). The authors will estimate the influence of response-related parameters on the masked priming effects and study the neural underpinnings of our dual-tasking manipulations. For that, response modality (vocal or motor) and task complexity (low vs. high complexity) will be manipulated, and how these two factors affect masked priming effects (i.e., incongruent trials – congruent trials) and the P3b component of the ERPs will be studied.

Overall, the proposal is interesting, as both methodological caveats of the priming paradigm and the cognitive correlates of the P3b are hotly debated topics. There are some issues, however, which I believe the authors should address before a recommendation can be made on the manuscript.

Introduction:

- Pages 3-4. The authors might want to consider the recent study by Jimenez et al. (2023) when presenting the studies that have used single and dual task priming designs. In this study, the authors discuss the dual-task character of the designs when indirect and direct tasks are presented together. Their results showed an increase in overall RTs in the dual-task condition as opposed to the single-task condition, where priming effects were found at specific prime-mask SOAs and an overall decrease in RTs was observed.

We added Jimenez et al., 2023 as a reference for studies comparing single and dual-task conditions and finding longer RTs for dual as compared to single-task conditions. However, Jimenez and colleagues did not find masked priming effects in experiments 1-3 in their dual-task condition. Only when the prime was not masked (experiment 4) did priming effects show in dual-task.

- Page 5, first paragraph. The study by Biafora & Schmidt (2022) is not explained in the Intro. Since it seems important to the current study, the reader might benefit from a brief description of that study.

This is a good point, and we added the following description to the introduction: "Indirect and direct task have been presented together (e.g. Stein et al., 2021) as well as in separate trials (e.g. Biafora & Schmidt, 2019). Biafora and Schmidt (2022) combined both approaches and compared a single-task condition (either only indirect or direct task) with a dual-task condition, for which they instructed participants to perform both a target (mask) identification task and a prime identification task on the same trial (experiment 2). The authors observed increased RTs and larger priming effects in the dual-task as compared to the single-task condition." (p. 4)

- Page 5, second paragraph. "One commonly used experimental design in the line of masked (unconscious) priming research is metacontrast masking (e.g. Mattler, 2003; Vorberg et al., 2003)." The authors might want to consider including additional references, such as the review by Breitmeyer (2015), for further insights on the different techniques to render a stimulus invisible.

We included Breitmeyer, 2015 in the reference here. (p. 7)

- Page 6. The aim of the section presented here (Metacontrast-masked response priming and Dual-tasking) is not very clear. Do the authors want to explain that meta-contrast masking is especially suitable to assess priming effects in dual-task paradigms? On the other hand, how does metacontrast masking specifically relate to the PRP and BCE phenomena?

We are not trying to argue that metacontrast masking is especially suitable to assess priming effects in dual-task paradigms. We changed the section heading to "Masked Priming and Dual-Tasking". (p. 6) Metacontrast Masking is just our choice of method, and masked priming paradigms gain a dual-task characteristic when the indirect and direct tasks are applied in the same trial.

- Page 6, last paragraph. It is a bit difficult to understand the experimental design of Scerra and Brill (2012). Authors may want to consider rephrasing, for example: " Scerra and Brill (2012) tested participants in several multitasking experiments, in which the input of both tasks was either presented in the same modality (visual prime and target; unimodal dual-task condition) or via different modalities (tactile prime and visual target or tactile prime and auditory target; cross modal dual-task condition)."

That's a good point, thank you. We changed the manuscript accordingly. (p. 8)

- Page 9. The authors use Task 1 (probe response) and Task 2 (prime response) nomenclature. Later in the manuscript (e.g., age 19) the authors use 'indirect task' and 'direct task' instead of task 1 and task 2. I will advise for a consistent naming through the manuscript.

We agree that this should be consistent, and we have revised the manuscript accordingly. When other studies are described, the manuscript still uses task 1 and task 2, because not all used direct and indirect tasks.

- Page 10, second paragraph. Further references on objective and subjective measures of awareness might be added to the one by Hesselmann (2013). A recent review on the different measures of awareness can be found in Jimenez et al. (2024) which the reader might find interesting. Also, a more in-depth discussion can be found in Overgaard, 2015; an easier read on Persuh, 2018.

We added both Jimenez et al., 2023 and Overgaard, 2015 to our references here. (p. 4)

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

Overgaard, M. (2015). *Behavioural Methods in Consciousness Research*. Oxford University Press.-

Page 11, first paragraph. It will probably suffice to say that the PAS instructions were administered in German.

We removed the German translations and only mentioned that they were translated into German. (p. 13)

Page 12, second paragraph. A recent review on the P3b by Verleger (2020) might be added as a reference.

Done. Thank you for the suggestion.

Methods:

- Page 14, last paragraph. Block F will be assessed in a separate session without an EEG recording. Will Block F (prime identification task) be administered in the same day? The measurement of prime awareness would be ideally performed just after the Block E (single task).

If time and participants' patience allow, we will administer block F the same day. If that is not the case, we will ask participants to return for block F. (p. 17)

- Page 15. A single SOA of 94 ms (plus 24 ms prime presentation) will be used in the experiment. How is the SOA duration determined and justified, is it based on previous research? Will this stimulus presentation ensure PAS reports in all 4 (or 2, depending on the condition) categories?

Please see our response to reviewer M. Kiefer's comment regarding this issue:
We re-analysed the data from a pilot experiment where we used 2, 5, and 8 frames. Indeed, the SOA of 8 frames was associated with rather high PAS levels (mean PAS: 1.66; 0-3 PAS). We therefore decided to use a 5 frame SOA instead (appr. 59 ms) as it more adequately complies with our requirements of low visibility (mean PAS: 1.06; 0-3 PAS).

- Page 15. I wonder how the authors will assess that the participants are correctly using the PAS. This is normally evaluated by introducing catch-trials (e.g., prime absent trials).

We initially planned to omit catch trials to save on time, but we agree that it would be reasonable to assure that the participants are using the PAS correctly. We therefore plan to include catch trials and added following section to the manuscript:
"We will employ 5 catch-trials, i.e., trials missing a prime, to ensure correct use of the PAS: high visibility ratings in response to a catch-trial will tell us that the participants either did not pay attention or chose the label randomly." (p. 19)

- Page 19. Regarding hypothesis 3, if the results show an absence of RT differences and priming effects between 2- and 4-point PAS, how will the authors interpret these results? In other words, is it possible that the PAS response manipulation does not lead to increased task complexity?

This is a good point. Unfortunately, we do not have a manipulation check for our manipulation of task complexity. Therefore, we will need to address this issue in the discussion section of the final manuscript. Following McDowd and Craik (1988), who defined the increase in complexity as "associated with a greater degree of choice" (see paragraph "Task Complexity"), we believe that our manipulation will be successful.

- Page 19. Since the authors will explicitly test RT differences between conditions, it may be convenient to explaining why the 1.5 interquartile range (IQR) will be used here, and this is preferred as opposed to alternative approaches.

Thank you for bringing this to our attention. Based on the reviewer's comment, we have reconsidered our choice of outlier detection and consulted the paper by Berger & Kiefer (2021). Methods excluding RTs as outliers based on z-scores (e.g., "2sd" or "3sd") were associated with a low bias and clearly outperformed the absence of outlier exclusion in terms of bias. Bias was defined as the deviation of the proportion of significant differences after outlier exclusion from the proportion of significant differences in the uncontaminated samples (before introducing outliers). Therefore, we decided to use the "2sd" method for outlier detection and added the reference below.

Berger, A., & Kiefer, M. (2021). Comparison of Different Response Time Outlier Exclusion Methods: A Simulation Study. Frontiers in Psychology, 12, 675558. https://doi.org/10.3389/fpsyg.2021.675558

- Page 20, first paragraph. It is not clear whether in the dual-task condition, all trials will go to the analyses, or only trials for as specific PAS category (e.g., PAS1) will be analysed. If that's not the case, maybe these analyses can be included as supplementary?

We clarified this in the indicated paragraph: "All correct trials will be included in the analyses, regardless of PAS category." (p. 24) With only one SOA we guess it unlikely that there will be enough ratings for each PAS category to conduct such an analysis.

- Page 20, Exploratory Analyses. How do authors intend to explore P3b latency? Will the authors use a single-participant approach or the jackknife approach? Also, based on which specific method (e.g., peak latency, absolute criterion, relative criterion, fractional area) will the latencies be calculated?

Thank you for pointing this out, as this is indeed something we should have specified right away. Please see our response to Reviewer 2:

“We will be using the jackknifed averages for the mean amplitude and will follow a recommendation by Kiesel et al., (2008) to combine jackknifing with the relative criterion technique with parameter 50% for calculating onset latencies.” (p. 25).

- Page 21. Participants will report on their awareness of the primes. However, it is not clear whether the authors pretend to explore the unconscious processing of the primes or not. In the case of the dual-task blocks, that would involve including participants awareness (PAS) into the analyses, or otherwise exploring congruency effects for the PAS1 category.

This has indeed not been made clear enough yet in the manuscript. Thank you for pointing this issue out. We added the following paragraph to our *Design* section, in order to explain our use of the specific SOA further: “. Please note that we will use only a single SOA due to time constraints. Wean SOA We are primarily interested in a paradigm of low or reduced visibility (Handschack et al., 2022, 2023), since the main purpose of this study is to investigate the influence of a dual-tasking structure and that of manipulations of task 2 on RTs and priming effects.” (p. 18)