I now have detailed reviews back from two experts, both of which have made very useful points. They raise issues of, amongst other things, consistency within the document regarding analyses and predictions; clarity of stopping rule; possible other outcome neutral tests and other checks (e.g. regarding phosphenes as a measure of drift); and clarity concerning the TMS procedure.

On the analysis, I am in favour of planned comparisons, as in a Registered Report one wants to remove not only all analytic flexibility but also inferential flexibility in interpretting the analyses, so that the use of auxiliary assumptions is not biased - and this is made much easier by planned comparisons. Nonetheless, both reviewers queried whether you wished to make no comment about the diference between timings. I am not saying you should test such diferences, and explroatory analyses can go in a separate exploratory section; but the question is what contrasts directly test substantial theory, as those contrasts will inform e.g. the main conclusions in the abstract (and the discussion). The same point about what can go in exploratory analyses and therfore not foregrounded in conclusions applies to e.g. analysis of RTs. Just bear in mind in responding to reviewers on this issue, in a Registered Report, to keep things clean, one does not pre-register exploratory analyses.

In terms of motivating your scale factors for the Bayes factors, I do not understand the main motivating sentence "The width parameter of each prior was calculated to correspond to the 90% probability of the effect size lying within the standardised differences (Hedge's g) of accuracies and signal detection estimates between sensory visual cortex TMS and a control condition reported in a recent meta-analysis on the topic (Phylactou et al., 2021)." Do you mean you used a 90% CI on the diference scores? Or a 10th percentile of the diferences distribution? Bear in mind that a Bayes factor requires a roughly expected effect size for the scale factor (not e.g. a minimal meaningful effect); the meta-analytic effect sizes you quote are large compared to your scale factors - why not just use relevant meta-analytic effects? Additionally, you might consider reporting a "Robustness Region" for each BF to show how robust it is to changes in scale factors (as defined here https://psyarxiv.com/yqaj4); I leave this final point up to you.

Thank you very much for all the useful feedback and support throughout this process. We express our apologies for the delayed resubmission, which was due to unexpected drawbacks from changes on the national Covid-19 protocols. Below, you can find a point-to-point response to the reviewers' concerns and suggestions.

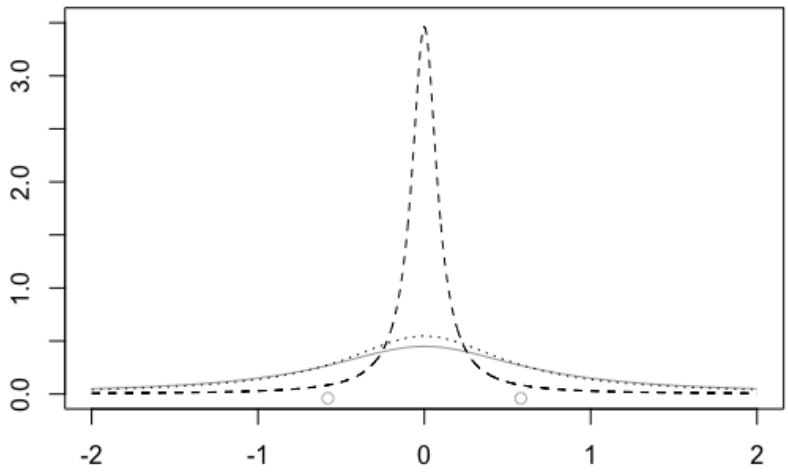
We would like to mention that we have now updated our priors following new results in the meta-analysis that we had followed to inform our priors (Phylactou et al., 2021). On p.20, first paragraph, we now describe the new expected effect sizes: "By considering the overall effect size (g = .58), the effect size for early TMS (up to 200 ms; g = .80), and the effect size for late TMS (after 200 ms; g = .50) from previous meta-analytic work (Phylactou et al., 2021), the width parameter of the Cauchy distribution was calculated to correspond to 0.092 for the 0 ms condition, to 0.13 for the 200 ms condition, and to 0.08 for the 1000 ms condition, respectively."

Regarding our scale factors, the percentage (90%) corresponds to the area under the curve of the Cauchy distribution. To define the priors, we calculated a Cauchy distribution, where all values between the relevant effect size from the meta-analysis lay within 90% of the distribution. The rationale behind this is driven by the approach used to calculate the default prior used in Jamovi (Jamovi, 2021). The default prior is a Cauchy distribution centered around 0, with a width 0.707 (Morey & Rouder, 2011; Rouder et al., 2009). This prior creates a distribution where 80% of the distribution contains an effect size between -2 and 2. As such, adjusting the width parameter of this default in Jamovi, still corresponds to testing for a true effect size approximately -2 and 2. Therefore, to define our priors, we calculated the Cauchy distributions which correspond to a distribution where 90% of its values lay between effect sizes -X and X, where X equals to the effect sizes reported in the meta-analysis. This was done by adjusting the width of a Cauchy distribution centered on 0 at locations -X and X, until the area under the curve (or cumulative probability) was as close to 90% as possible. We chose the width that led to a 90% cumulative probability

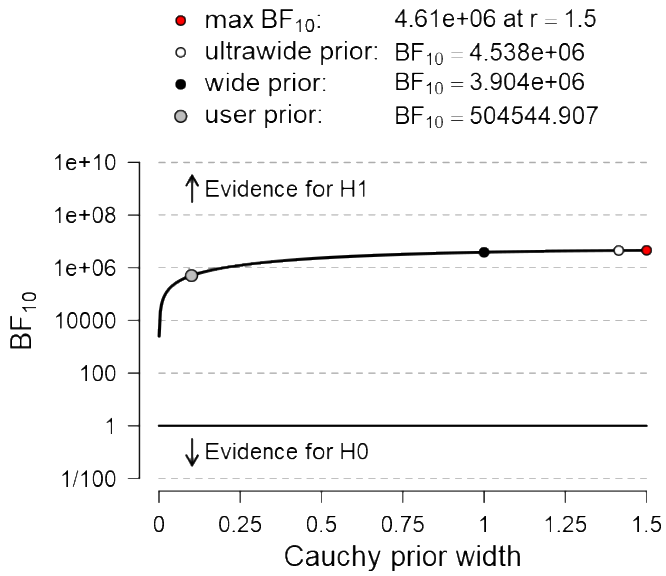
of containing the values -X and X over (the 'default') 80% to reduce Type I error. Given so, the reason our scale factors are smaller than the effect size is because we used the meta-analytic effect sizes as the locations of the Cauchy prior. Put simply, our priors assume that the null distribution is centered around 0 with a 90% cumulative probability of containing values between the locations -X and X, where X = expected effect size.

We have rephrased this in the revised version in order to clarify, as such (p.20, first paragraph): "Each prior for the paired t-tests is described by a Cauchy distribution centered around zero (see Rouder et al., 2009). Each prior was based on the results of a recent meta-analysis on the topic (Phylactou et al., 2021), which reported the standardised differences (Hedge's g) of accuracies and signal detection estimates between sensory visual cortex TMS and control conditions. The width parameter of each Cauchy prior was calculated to correspond to a 90% cumulative probability of values in the distribution laying within the expected effect size." An alternative in case this is not appropriate would be to simply use the expected effect size as the width of the prior, where *r = g* (i.e., the reported effect size in the meta-analysis).

Below, we present a graph we prepared in R, showing the densities (y-axis) and effect sizes (x-axis) of three Cauchy distributions to illustrate the example described above. For illustration purposes, we have also plotted values -0.58 and 0.58 (the revised expected effect size for our 0 ms condition), which are presented as grey open circles. The grey line represents the JAMOVI default prior with width r = 0.707. The black dotted line is the Cauchy distribution with the width set as the expected effect size (r = 0.58). The black dashed line is the distribution we have calculated, where the cumulative probability of the expected effect size (g = 0.58) is 90%, thus resulting in a width r = 0.092. As one can see from the graph, our prior (black dashed line, r = 0.08) assumes that effect sizes greater than 0.58 or smaller than -0.58, are less likely, compared to the 'default' Cauchy prior (grey line, r = 0.707) or a prior with width equal to the expected effect size (black dotted line, r = 0.58), which both allow for greater likelihoods for values > 0.58 and < -0.58.



As far as the robustness region, our analyses will include BF robustness check analyses. The robustness check will graphically illustrate the value of the BF according to the width of the Cauchy prior, for widths from 0 to 1.5. This analysis will also indicate the r which produces the maximum possible BF, as well as the BF for a wide (r = 1) and ultrawide (r = 1.4) prior. Below, you can see an example of such a graph generated from a study we are currently conducting. However, we considered the robustness check analyses as part of our exploratory analyses and this is why they are not discussed in the current version of the registered report.



Reviews

Reviewed by Evie Vergauwe, 16 Nov 2021 08:39

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

The proposed study is concerned with the role of the sensory cortex is visual short-term memory. In particular, the study aims to test whether the sensory cortex plays a causal role in short-term maintenance of visual information, which is a highly relevant question that follows directly from the existing scientific literature and evidence base. The question is defined with sufficient precision to be answered through appropriate experiments.

Thank you for the feedback and the positive evaluation of our work.

1B. The logic, rationale, and plausibility of the proposed hypotheses (where a submission proposes hypotheses)

The study aims at testing a clearly-defined hypothesis: that of whether the visual cortex plays a causal role in visual short-term maintenance. If that is the case, then interfering with the activity of the visual cortex during short-term maintenance should have an impact on visual short-term memory performance. The hypothesis follows directly from the research question and the currently-available theory in the field. The hypothesis is very popular in the field, and much recent research has aimed at testing it, either directly or indirectly. However, there seems to be some inconsistency between the text and Table 1 when it comes to the exact description of the hypotheses: 1) the text seems to argue that the authors aim at testing that there will be *no* difference between the ipsilateral and contralateral conditions", see p. 8, line 7 from the bottom, even when TMS is applied during perceptual processing (i.e., the outcome neutral condition), whereas the table states "Given the established role of the sensory visual cortex during visual perception, we hypothesize that evidence *for* a difference between the ipsilateral and contralateral conditions will be present when sensory visual cortex TMS is induced at 0 ms". 2) in the text, it is explained on p. 10, that, for the outcome-neutral condition "a significant drop in VSTM performance (decreased detection sensitivity) is expected in the ipsilateral compared to the contralateral condition)." This is a hypothesis with a clear direction: decreased detection sensitivity. However, in Table 1, the corresponding t-test appears to be two-sided, and it

appears that "any" difference between ipsilateral and contralateral conditions would be considered as evidence for the hypothesis, with sample means used to indicate "whether the effects of TMS are inhibitory (sample mean < 0) or facilitatory (sample mean > 0)." Finally, one thing that was not considered is the idea that information in visual working memory can be represented in a continuous and a categorical way – it is not entirely clear whether, at the theoretical level, the authors aim specifically at the continuous memory representations or at both.

Thank you for pointing out the inconsistency regarding our hypotheses in the text compared to Table 1. We have revised the main text to match the hypotheses in Table 1, as based on the current state-of-the-art we believe that we cannot confidently hypothesize on the direction of the TMS effects. This is mainly due to the fact that previous studies presented the memory sample binocularly, thus it is possible that the direction of effect is falsely interpreted (or undetectable). This is described in the registered report on p.7 paragraph 1:

"As with different methodological approaches, results from previous TMS studies were mixed with regards to the sensory recruitment hypothesis. Specifically, some of the studies supported the sensory recruitment hypothesis (Cattaneo et al., 2009; Jia et al., 2021; Silvanto & Cattaneo, 2010), some rejected it (Rademaker et al., 2017; van Lamsweerde & Johnson, 2017), while others were unclear (van de Ven et al., 2012). After a careful examination of the methods used in previous TMS studies, we suggest that the inconclusive findings are due to several important methodological issues that may have underestimated the contribution of the sensory visual cortex in VSTM. The most vital issue in the majority of these TMS studies is that previous researchers considered that, when information was presented on one side of the visual hemifield (either right or left side near the centre of the monitor), the information was processed by the contralateral sensory visual cortex. Therefore, stimuli were always presented binocularly to the participants either in the left or right visual field, and a contralateral sensory visual cortex TMS was applied and compared to an ipsilateral control condition (see Figure 1A). However, considering the neuroanatomy of the visual pathway system, the binocular presentation of stimuli either left or right close to the midline of the visual field -as was the case in the majority of the previous studies- does not accurately correspond to the contralateral sensory visual cortex, and could in fact be processed by the ipsilateral one if presented withing 15^{o} of visual angle from midline (Joukal, 2017; Wichmann & Müller-Forell, 2004). It is also possible that information enters the sensory visual cortex in both brain's hemispheres (Tong et al., 2006; Zhao et al., 2021) since the visual fields of both eyes overlap in certain areas (within 15° of visual angle, see Figure 1B; Wichmann & Müller-Forell, 2004). Consequently, some TMS effects can be falsely interpreted or remain undetectable (e.g., if information processing happens in both hemispheres despite the contralateral and ipsilateral conditions; de Graaf & Sack, 2011; see also Pitcher et al., 2020). For example, as pointed out in a recent review (Phylactou et al., 2021), a study considering the contralateral TMS condition as the experimental condition and the ipsilateral side as the control condition will interpret a performance drop (e.g., contralateral performance < ipsilateral performance) as an inhibitory TMS effect. Though, considering the evidence supporting the role of the ipsilateral sensory visual cortex in visual processing (Zhao et al., 2021) and the neuroanatomy of the visual pathway (Wichmann & Müller-Forell, 2004), it is possible that the ipsilateral sensory visual cortex is in reality the experimental condition. As such, the conclusion of the same study, might turn out to be the opposite (e.g., facilitation effects since ipsilateral accuracy > contralateral accuracy), if the experimental and control conditions are inversely defined."

Further, to address the inconsistency, we rephrased the main text to correspond to the hypotheses in Table 1 as indicated below.

P.10, last paragraph:

"To explore our main question of whether the sensory visual cortex is involved in VSTM maintenance, our hypotheses focus on testing differences in detection sensitivity (Stanislaw & Todorov, 1999) in a VSTM task, between the ipsilateral and contralateral conditions when stimuli are presented monocularly and

TMS is applied (1) during perceptual processing (outcome neutral condition; 0 ms after stimulus onset), (2) during early information maintenance (200 ms after stimulus onset), or (3) during late information maintenance (1000 ms after stimulus onset)."

P12, last paragraph:

"Thus, given the established role of the sensory visual cortex during visual perception (D'Esposito & Postle, 2015; de Graaf et al., 2014; Kamme, 2007; Serences, 2016, Xu, 2017), a significant difference in VSTM performance is expected in the ipsilateral compared to the contralateral condition (either facilitation or inhibition; for details see Table 1)."

We share the theoretical interest for the categorical and continuous representations in VSTM. Indeed, recent work has pointed out that representations in VSTM can be stored in both categorical and continuous ways (e.g., Hardman et al., 2017 for colors, Bae, 2021 for orientations and spatial locations). However, given our key questions of interest, we decided to not introduce this topic in the current Stage 1 of our registered report for the following reasons:

1. Recent psychophysical work has shown that categorical judgments in an orientation VSTM task do not modify sensory representations (Luu & Stocker, 2021). We, therefore, assume that sensory visual cortex TMS is likely to interfere with VSTM representations regardless of whether representations are stored categorically or continuously.
2. Here we are using two-alternative forced choice (2AFC) orientation change detection tasks with responses limited to either a correct or an incorrect response. As such, the data restrict us from fitting computational models as those implemented with delayed-estimation tasks in order to quantify categorical and/or continuous representations in VSTM (e.g., Bae, 2021; Hardman et al., 2017; Luu & Stocker, 2021). The choice of a 2AFC task over a delayed estimation task, stems from our main research question, which is whether the sensory visual cortex is necessary (or not) for VSTM maintenance. The current debate relates to the involvement of the sensory visual cortex, and so we find the 2AFC more appropriate to answer the main question of whether TMS affects VSTM performance. Further, to avoid inferences based on the accuracy results of the 2AFC tasks (i.e., using a percentage variable for statistical tests of continuous variables, such as *t*-tests), we will draw our main conclusions from the standardized detection sensitivity index *d'* (see *Analysis Plan,* p17). Unfortunately, *d'* is not informative of whether the representations were stored as categorical or continuous, however it can inform us of the sensitivity to discriminate between change and no change, which is the main variable of interest for our hypotheses.

However, during Stage 2 of the proposed work, we aim to discuss in some extend the topic of categorical and continuous representations. For example, the latest theoretical framework describing a distributed nature of sensory VSTM (Lorenc & Sreenivasan, 2021; Teng & Postle, 2021), postulates that the sensory visual cortex can be flexibly involved during VSTM maintenance, depending on available resources and task demands. One such example is the use of different working memory processes, such as verbal memory, when visual stimuli have been verbalized (e.g., Overkott & Souza, 2021), which might result in the representations being stored in higher order brain areas, other than the sensory visual cortex. Moreover, categorical representations in VSTM have been alleged as a possible byproduct of verbalizing stimuli (Hardman et al., 2017; Souza et al., 2021). We think that this line of work can guide future research in order to gather empirical evidence for current theoretical propositions (e.g., as the ones described in Lorenc & Sreenivasan, 2021 and Teng & Postle, 2021). Therefore, in the discussion section of the Stage 2 phase of our work we will be describing the importance of focusing future work on the distinction between categorical and continuous representations.

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

Overall, the methodology and proposed analysis pipeline appear as feasible and sound. I am not an expert in TMS so I cannot speak to the technical details of the proposed TMS procedure. 1) As a general comment on the proposed method, the authors argue, in the introduction, that the inconclusive findings in the literature are due to several important methodological issues with these studies and pinpoint two of them: binocular presentation (as opposed to monocular presentation) and complex stimuli (as opposed to simple stimuli). However, this point is made rather superficially; it would be more convincing to explain more in detail how these issues could account for the contradictory findings of study x vs. study y. This, in turn, would support the proposed method as the best way to address the issue, to obtain a conclusive answer to the question. 2) Bayesian sequential hypothesis testing is proposed, which is appropriate. However, the exact wording could be more precise as to avoid any ambiguity. For example, it is explained "Sample updating with a stopping rule set at BF10 > 6 or < 1/6 for all three paired t-tests." – it may be useful to explain in detail that this means that, if one of the three t-tests does not reach the predefined criterion, testing will continue (if I understood correctly). Also, a minimum of 20 participants and a maximum of 40 participants is mentioned, but it is not entirely clear whether this implies that after a first batch of 20 participants, a second batch of 20 participants will be tested if needed (as opposed to, for example, adding participants in small batches of 5 after the initial batch of 20 participants). 3) As far as data exclusions are concerned, not much detail is provided. Only participants with less-than-optimal color vision are planned to be excluded. It is not clear what will happen in case of technical difficulties leading to loss of part of the data of a given participant. No performance-based exclusions are proposed. However, one could expect that only participants with a certain level of memory performance (better than guessing) would be included. For participants who are not really doing the task (i.e., are not trying to remember the orientation), it is difficult to expect that interfering with the visual cortex will impact their performance.

Thank you for sharing these concerns regarding our methods.

1) We thank the reviewer for suggesting an improvement for the rationale behind the methodological issues with previous studies. Regarding binocular presentation, the rationale is based on the neuroanatomy of the visual pathway, which we believe was not accounted for in previous studies. This was also reported in our recent systematic review and meta-analysis (Phylactou et al., 2021), where the review of studies comparing the ipsilateral to the contralateral conditions, indicated an issue regarding the reliability of their results. In the revised registered report, we have elaborated on the rationale for the use of monocular stimulus presentation, as presented on p.7 paragraph 1 and outlined in response to your comment 1B above.

With regards to the use of complex stimuli, we have now rephrased the paragraph accordingly and provided additional relevant work. On p.8, first paragraph, we now mention:

"Another important shortcoming relates to the complexity of the stimuli used in the memory array. In a given memory array, there is a minimal representational requirement for VSTM, based on the core features (e.g., color, orientation, shape) of stimuli (Alvarez & Cavanagh, 2004). A greater combination of stimuli features increases complexity and VSTM capacity requirements. Previous TMS studies used various stimuli in their memory tasks, some of which were complex stimuli such as abstract shapes (van de Ven et al., 2012). However, the evidence leading to the sensory recruitment hypothesis emphasized the selective engagement of the sensory visual cortex in elemental visual features such as orientation, contrast, and direction of movement (Harrison & Tong, 2009; Issa et al., 2008; Konstantinou et al., 2012; Serences et al., 2009). For example, Jia and colleagues (2021), indeed found a strong TMS effect in a VSTM task requiring participants to remember the elemental visual feature of orientation of one grating, but in a study requiring participants to remember either one (low load) or three (high load) abstract shapes (that are thought to be complex stimuli consisting of a combination of elemental visual features; van de Ven et al., 2012), TMS did not affect performance in the low load condition of remembering a complex shape (TMS effects were only evident during the three-item high load condition). Such findings are suggesting that when stimuli complexity increases, higher order brain areas might be more actively recruited for VSTM, such as the intraparietal sulcus (Xu & Chun, 2006; Xu, 2007) and the posterior

parietal cortex (Song & Jiang, 2006). Thus, the neural processes required for successfully maintaining complex visual stimuli in VSTM might be more dependent on higher order brain areas than simple stimuli consisting of elemental visual features, given the high selectivity of sensory visual cortex in processing of elemental features (Teng & Postle, 2021). This might explain some of the null effects of sensory visual cortex TMS during the memory delay, since complex representations are likely protected through a more distributed VSTM network (Lorenc & Sreenivasan, 2021; see also Gayet et al., 2018; Scimeca et al., 2018). Hence, it is possible that some of the previous studies failed to find evidence for the sensory visual cortex involvement in VSTM due to using such more complex stimuli."

2) Following the suggestion of the reviewer, we have now expanded our sampling plan as follows (p.18, second paragraph):

"For Experiment 1, sample updating with a stopping rule has been set to BF10 > 6 or < 1/6 for all three paired t-tests that will be performed. However, due to counterbalancing, a minimum of 20 participants will be gathered (to ensure counterbalancing) or a maximum of 40 participants, given time and resource constraints. Specifically, after data collection for the first 20 participants is completed, we will perform our analyses to check if the stopping rule has been fulfilled. If any of the three BFs did not reach the stopping rule of > 6 or < 1/6, we will recruit four more participants and repeat the analyses. This process will be repeated until all three BFs fulfil the stopping rule, or until the maximum of 40 participants is reached."

3) We thank the reviewer for pointing out the omission regarding data exclusion details, and we apologise for this oversight. In the revised version, we have added a *Data Filtering* section (p.21, *Data filtering* section), where we describe the following:

"*Data filtering. Participants with an overall accuracy close to chance levels (< 60% accuracy) in Experiments 1 and 2 will be excluded from analyses and replaced. The data of such participants will not be used during Bayesian sample updating nor for our main analyses. Additionally, we will exclude and replace participants in the case of technical or other difficulties, if data loss is greater than 20% of the total experimental trials. Further, the slowest and fastest responses will be removed from the analyses. To do so, we will filter each participant's responses and exclude any data that concern response times which are further than 3 standard deviations away from each participant's mean reaction time. Assuming that the reaction times of each participant are normally distributed, we expect less than 0.5% of the data of each participant to be excluded from the main analyses.*"

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

The methodology and proposed analyses are described sufficiently in detail to replicate and to prevent undisclosed flexibility. One thing to note, though, is that the introduction mentions "We propose to test the hypothesis that, when visual stimuli are presented monocularly, participants' average detection sensitivity (Stanislaw & Todorov, 1999), accuracy, and response time in a delayed change-detection VSTM task will not differ between the ipsilateral and contralateral conditions when TMS is applied", whereas the proposed analysis plan appears to only use detection sensitivity as outcome variable. Will accuracy and reaction times also be analyzed? And if so, how will these findings be interpreted in the light of the findings on d'? Another point is that the authors are only proposing t-tests (e.g., one per timing condition) and no statistical test of any interaction. They will need to be very careful to interpret their results as to not make any claims that would have required testing the corresponding interaction (such as stating that the effect of TMS depended on when it was applied, which would require testing the interaction of Presentation side x Time of TMS).

These are very relevant concerns, and we thank the reviewer for pointing them out, allowing us to clarify that our analysis plan is limited to *t*-tests on *d'*. Since this is a Stage 1 RR, we restricted our analysis plan only to the most relevant statistical tests that concern our main research question (*is the sensory visual cortex involved during VSTM maintenance*) and did not describe exploratory analyses, in order to keep the registered report as clean and focused on the main hypotheses. We plan to perform and report statistical tests (i.e., repeated measures ANOVA) to explore the interactions between TMS timing (0 ms, 200 ms, 1000 ms) * TMS site (ipsilateral vs contralateral) in Experiment 1 and TMS timing (200 ms, 1000 ms) * TMS site (ipsilateral vs contralateral) * TMS condition (real vs sham) in Experiment 2 for *d'*, however we consider these tests as part of our exploratory analyses since they do not directly test our main hypothesis. Further, we will also perform and report exploratory analyses on reaction times, which will be similar to the analyses performed on *d'* (repeated measures ANOVA and t-tests). All analyses described above will be included in an *Exploratory Analyses* section during Stage 2 of our registered report. Accuracy will not be analysed, since this is a percentage variable, but is often mistakenly treated as a continuous variable in the field (e.g., by performing a t-test or ANOVA, which assume that data are continuous). One way of overcoming this statistical error is by arcsine transforming accuracies (e.g., Lin & Xu, 2020), or standardizing them, by transforming them to continuous measures, such as *d'*, which is considered more appropriate (e.g., Stanislaw & Todorov, 1999). Since we will perform analyses on *d'*, we find further analyses on accuracy to be redundant.

We also thank you for identifying the inconsistency in the phrase pointed above "*We propose to test the hypothesis that, when visual stimuli are presented monocularly, participants' average detection sensitivity (Stanislaw & Todorov, 1999), accuracy, and response time in a delayed change-detection VSTM task will not differ between the ipsilateral and contralateral conditions when TMS is applied*". This has been amended in the revised registered report (p.10, last paragraph):

"To explore our main question of whether the sensory visual cortex is involved in VSTM maintenance, our hypotheses focus on testing differences in detection sensitivity (Stanislaw & Todorov, 1999) in a VSTM task, between the ipsilateral and contralateral conditions when stimuli are presented monocularly […]".

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).

One outcome-neutral condition is proposed in Experiment 1: a difference in memory performance between ipsilateral and contralateral conditions when TMS is applied during encoding/perception (as opposed to TMS during retention). This is a strong positive control. As pointed out in Table 1, in case no difference between ipsilateral and contralateral conditions would be found in this control condition, it would indicate that TMS effects are undetectable between the ipsilateral and contralateral conditions. In that case, I think it would be useful to also have an outcome-neutral condition in Experiment 2, for the other proposed comparison, whereby one would test whether there is a difference in memory performance between real vs sham TMS when TMS is applied during encoding/perception (as opposed to TMS during retention). Also, there is currently no behavioral quality check; participants should reach a minimum level of memory performance to be included, in my opinion.

We thank the reviewer for raising these points.

Regarding the positive control in experiment 2, as described on p.18, first paragraph, the 0 ms condition was dropped due to the use of the sham TMS condition. Sham TMS can be compared to a "placebo" condition. As such, we will have two control conditions in Experiment 2: (1) sham TMS versus real TMS, and (2) ipsilateral TMS condition versus contralateral TMS condition. Having two control conditions is considered very reliable when exploring TMS effects (Deuker & Sack, 2015; Pitcher et al., 2020). Similarly, in Experiment 1, we deem the outcome neutral condition necessary, because stimulation is always present, and it can inform us whether TMS effects can indeed be quantified between the ipsilateral and

contralateral conditions. This confirmation is not necessary for Experiment 2, since we can directly compare performance between real TMS (i.e., actual stimulation) and sham TMS (i.e., no stimulation).

Including an additional 0 ms condition as a positive control is likely to result in a very long task for the participants (700 < trials), thus risking the quality and reliability of the collected data. Therefore, we believe that including a third control (e.g., an outcome neutral condition) in Experiment 2, might not be necessary.

Regarding the behavioral quality check, in addition to the proposed staircase procedure, which will estimate the individual orientation difference for discriminating with an approximate accuracy of 75%, we will also ensure a 75% accuracy during practice trials, before proceeding with the main experiments. We have added this information on p.17, last paragraph starting on p.16:

"Furthermore, before the two main experiments, participants will carry out a practice block of 24 trials without TMS stimulation to familiarize themselves with the experimental procedure. If accuracy in the practice block is less than 75%, the practice block will be repeated until the participant reaches at least 75% accuracy. Participants will be replaced if after four practice blocks their accuracy remains below 75%."

Also, and further to the third comment in section 1C above, we would also remove outlier trials (based on slow or fast responses) and replace participants with performance close to chance levels (<60% overall accuracy), as we now describe in *Data filtering* section of the registered report (p.21).

Reviewed by Vincent van de Ven, 08 Dec 2021 16:07

I reviewed the registered report by Phylactou et al., in which the authors propose a TMS study to investigate the contribution of lateralized (early) sensory cortex in the maintenance of visual information in short-term memory. The authors propose to use a novel combination of occipital TMS with dichoptic visual presentation that allows presentation of visual information to only one eye, such that it is processed in one hemisphere. The authors plan to conduct two experiments, in which sham TMS is to be used in experiment 2.

1) Overall, I think the motivation and argumentation for the study are strong and well developed. The coverage of literature and consideration of current shortcomings in TMS studies of the role of visual cortex is thorough and elaborate. The authors make a clear case about why another TMS study of the role of sensory cortex in short-term memory maintenance is needed. However, I feel that the authors should acknowledge and discuss the current debate in the literature between Xu and others about whether current evidence indeed supports a memory-maintenance role of sensory cortex vs. a more perceptual/encoding/attentional role. To be sure, the authors cite the relevant works and published opinion pieces, but I feel the discussion is rather left untouched – given the focus of the proposed experiments, this issue must be considered at one point (here or in manuscript of the results), to understand how evidence would eventually weigh into the debate.

Thank you very much for the valuable feedback. We agree with the suggestion that some elaboration on the debate could be valuable to the work. On p.5, we have added the following:

"Specifically, Xu (2017, 2018, 2020, 2021) argued that given the essential role of the sensory visual cortex in processing of incoming visual information (see Awh & Jonides, 2001; D'Esposito & Postle, 2015; de Graaf et al., 2014; Kammer, 2007; Masse et al., 2020; Serences, 2016), it is unlikely that it is also involved in visual information maintenance since such dual processing by sensory visual cortex would result in maintained representations being susceptible to overwriting by incoming visual information. Other authors have pointed out that the sensory visual cortex protects representations by utilizing processes such as between layer top-down signals in area V1 (e.g., Gayet et al., 2018; Scimeca et al., 2018; van

Kerkoerle et al., 2017; Zhao et al., 2021). These processes are thought to be similar to those employed by the prefrontal cortex during attention modulation (see Scimeca et al., 2018), for example, when need to differentiate between mnemonic and perceptual information (e.g., Knight et al., 1999). Additionally, it has been proposed that the interaction between memory representations and perceptual input might be beneficial instead of detrimental to VSTM. For example, VSTM representations can improve perceptual continuity and goal-related behavior by biasing perceptual input (Gayet et al., 2013; Kiyonaga et al., 2017). Further, Xu (2017) pointed out that the sensory visual cortex is not sufficient to support sustained VSTM activity and sustained activity recorded in the sensory visual cortex most likely relates to feedback from higher order brain areas. However, alternative explanations proposed that sustained activity in the prefrontal cortex, might echo a biasing signal to protect or direct attention towards goal related VSTM representations, rather than reflect VSTM representations per se (Curtis & D'Esposito, 2003; Sreenivasan & D'Esposito, 2019; Masse et al., 2020; Miller & Cohen, 2001; Sreenivasan et al., 2014)."

Further, we anticipate that in Stage 2, a part of our discussion will focus on how our results can be interpreted in light of the current debate, as suggested by the reviewer.

A few more comments about the Introduction:

1) I find the paragraph (p7) about stimulus complexity in (not?) finding TMS effects rather unclear. If more complex stimuli tax memory more, it would not make it easier for TMS to impair processing? Or do the authors suggest that more complex stimuli are processed in higher levels of visual processing? What is meant with "elemental features" (complex stimuli also contain elemental features)? Please elaborate this notion or skip altogether – I would think the motivation to use gratings / gabors in this study does not depend on previous choices for complex stimuli.

Thank you for mentioning the need for elaboration on the notion of stimulus complexity. We have rephrased this paragraph (p.8, last paragraph), so that it is clear to the reader how null TMS effects might be expected when complex stimuli are used, as such:

"Another important shortcoming relates to the complexity of the stimuli used in the memory array. Previous TMS studies used various stimuli in their memory tasks, some of which included complex stimuli such as abstract shapes (van de Ven et al., 2012). However, the evidence leading to the sensory recruitment hypothesis emphasized the selective engagement of the sensory visual cortex in elemental visual features such as orientation, contrast, and direction of movement (Harrison & Tong, 2009; Issa et al., 2008; Konstantinou et al., 2012; Serences et al., 2009). For example, Jia and colleagues (2021), indeed found a strong TMS effect in a VSTM task requiring participants to remember the orientation of one grating (elemental visual feature), but in a study requiring participants to remember either one (low load) or three (high load) abstract shapes (complex stimuli; van de Ven et al., 2012), TMS did not affect performance in the one-item low load condition (TMS effects were only evident during the three-item high load condition). Such findings are suggesting that when stimuli complexity increases, higher order brain areas might be more actively recruited for VSTM, such as the intraparietal sulcus (Xu & Chun, 2006; Xu, 2007) and the posterior parietal cortex (Song & Jiang, 2006). Thus, the neural processes required for successfully maintaining complex visual stimuli in VSTM might be more dependent on higher order brain areas than simple stimuli consisting of elemental visual features, given the high selectivity of sensory visual cortex in processing of elemental features (Teng & Postle, 2021). This might explain some of the null effects of sensory visual cortex TMS during the memory delay, since complex representations are likely protected through a more distributed VSTM network (Lorenc & Sreenivasan, 2021; see also Gayet et al., 2018; Scimeca et al., 2018). Hence, it is possible that some of the previous studies failed to find evidence for the sensory visual cortex involvement in VSTM due to using such more complex stimuli.**"**

2) The stated hypotheses (p8) seems counterintuitive or otherwise unclear to me. What does "will not differ between ipsilateral and contralateral conditions" mean here? Surely, the authors do expect that

TMS at particular timepoints will affect processing in ipsilateral but not contralateral conditions? Further, what are the hypotheses about the timepoints?

In the main text, we have rephrased our hypotheses to reflect the descriptions presented in Table 1. The phrase "will not differ between ipsilateral and contralateral conditions" has now been revised and reads as follows (p.10, last paragraph):

"To explore our main question of whether the sensory visual cortex is involved in VSTM maintenance, our hypotheses focus on testing the differences of participants' average detection sensitivity (Stanislaw & Todorov, 1999) in a delayed change-detection VSTM task, between the ipsilateral and contralateral conditions when stimuli are presented monocularly and TMS is applied (1) during perceptual processing (outcome neutral condition; 0 ms after stimulus onset), (2) during early information maintenance (200 ms after stimulus onset), or (3) during late information maintenance (1000 ms after stimulus onset). Moreover, in a second experiment, we will test whether VSTM performance differs between a TMS and a sham TMS condition (1) during early information encoding (200 ms after stimulus onset) and (2) during memory maintenance (1000 ms after stimulus onset). Table 1 presents a detailed description of the research hypotheses for each experimental condition."

Regarding the differences between TMS timepoints, these will indeed be explored, but the relevant analyses will be reported in Stage 2, under our *Exploratory Analyses* section. The reason behind this is that, even though analyses between different timepoints are very relevant and interesting to the question regarding the involvement of sensory visual cortex in VSTM, they do not directly reflect our main question which is whether sensory visual cortex is involved during VSTM maintenance.

3) More generally, the role of the different timepoints in short-term consolidation, or the effect TMS could have on them is not discussed in the Intro at all. This must be spelled out as it is a key part of the TMS design in both experiments. What is the motivation to stimulate at 200ms after sample offset?

We apologize for omitting discussing the different timepoints in the introduction and we thank you for bringing this to our attention. We have included an additional paragraph in the revised report (p.9, last paragraph), elaborating on the rationale of our timepoints:

"Previous TMS studies, stimulated the sensory visual cortex at various timepoints during VSTM maintenance, with variable results (e.g., Rademaker et al., 2017; van de Ven et al., 2012; van Lamsweerde et al., 2017; for reviews see Phylactou et al, 2021; Xu, 2017). For example, Rademaker et al. (2017) interfered with sensory visual cortex TMS at 0 ms and 900 ms into the delay period, van Lamsweerde et al. (2017), at 0 ms, 100 ms, and 200 ms, and van de Ven et al. (2012) at 100 ms, 200 ms, and 400 ms. Some studies indicated that TMS effects were stronger for earlier stimulation (up to 200 ms; Rademaker et al., 2017; van Lamsweerde et al., 2017), compared to later stimulation at 400 ms (van de Ven et al. 2012), and 900 ms (Rademaker et al., 2017), however other studies indicated that TMS after 200 ms was stronger (van de Ven et al., 2012). Based on a recent meta-analysis examining the effects of TMS on VSTM performance during the maintenance period, most studies differentiated between earlier (up to 200 ms into the maintenance period) and later (after 200 ms; usually at the middle of the maintenance period) stimulation (Phylactou et al., 2021). In the current work, we similarly differentiated between early and late TMS, by considering the outcomes of previous studies (Rademaker et al., 2017; van de Ven et al., 2012; van Lamsweerde et al., 2017), in order to replicate and explore any similar findings concerning a different TMS effect size for earlier compared to later stimulation. Thus, we will examine the effects of TMS on behavioral performance separately for stimulation induced at 200 ms and 1000 ms (halfway) into the delay period."

Concerning the methodology, the general approach appears sound and is in keeping with previous designs, which facilitates comparison between studies. However, I do have several reservations and questions about the proposed design, which I will list below.

1) What is the motivation to use double-pulse TMS at 10 Hz?

We apologise for omitting the rationale for the choice of our TMS protocol. In the revised report, we have included the motivation behind double-pulse TMS at 10 Hz (p.11, first paragraph in *Apparatus and stimuli*):

"A 10 Hz double-pulse TMS was chosen to ensure the reliability of the outcome neutral condition. Specifically, the first pulse will be induced at the beginning of stimulus presentation and the second pulse at stimulus offset (see below). Given the possibility that a long encoding time (~100 ms) can lead to successful consolidation despite masking interference (Ye et al., 2017, 2021; Zhang & Luck, 2008), the double-pulse TMS will ensure that interference with regular brain activity is introduced throughout the consolidation process (Ye et al., 2017, 2021). For comparison and consistency reasons, the double-pulse TMS will be used in all experimental conditions."

2) The choice of sham TMS in experiment 2 is a valuable part of the experiment. However, the current manuscript provides very little explanation about sham TMS. The authors state that they will use an identical coil, but this seems incomplete or incorrect. If it is identical to real TMS, then stimulation procedures must be different (e.g., holding coil at angle to the head, or orienting the sham coil 90 degrees away from real TMS orientation). If procedures are identical, then a different coil must be used (e.g., thicker shielding to elicit acoustic and tactic sensation without inducing magnetic fields in the tissue). In short, information about sham procedures and coil must be further specified. Further, according to the authors, how likely will it be that participants will notice the differences between real and sham stimulation, and, if so, whether this could affect their findings? The pros and cons of sham stimulation are well described in a series of publications by Duecker and Sack (e.g., PLOS ONE, 2013; Frontiers in Psychology, 2015).

Thank you very much for bringing into the discussion the limitations of sham TMS, allowing us to improve the proposed work accordingly.

In order to elaborate on the use of sham TMS, we have rephrased our procedures as stated below (p.11, first paragraph in *Apparatus and stimuli*):

"A Magstim D70 Alpha Flat Coil (Uncoated) will deliver a double-pulse TMS at the different experimental conditions, while a sham coil will be used to control for noise and other TMS artefacts (in Experiment 2). The sham coil will look identical to the D70 Alpha Flat Coil, but it is equipped with thicker shield, restricting it from inducing magnetic fields, which interfere with brain activity."

Further, following the reviewer's suggestion, at the end of Experiment 2, which introduces the sham coil, we will ask participants to report whether they have noticed any differences between sham and actual stimulation. In the updated document, we report this on p.18, last paragraph before *Sampling Plan*, where we say:

"At the end of the experiment, participants will self-report whether they noticed any differences between sham TMS and TMS." This will allow us to proceed to further exploratory analyses in Stage 2, if differences in noticing sham TMS are reported.

We also share the concerns regarding some limitations of sham TMS, and TMS research in general. One consensus of the discussions relating to TMS research (e.g., Duecker & Sack, 2015; Pitcher et al., 2020) is the introduction of two control conditions, instead of relying on one control. This is something that we have considered ourselves, which is why in both experiments we have two control conditions. In Experiment 1, we compare the ipsilateral with the contralateral condition, and we also introduce the 0 ms condition which works as a positive control. In Experiment 2, we compare the ipsilateral to the contralateral condition, while also comparing sham to real TMS.

3) Some details about trial responses are missing. Will there be a limited response window? Will participants receive feedback about their responses? In the analysis, do authors plan to use data trimming strategies (e.g., discarding overly fast or slow responses) -- if so, please provide details?

We find these concerns very important and we thank you for pointing these out. Regarding the missing details about trial responses, we have updated the description of our experimental procedure (p.17 last paragraph), as such:

"Participants will have up to 3000 ms starting at probe onset to respond by placing their index and middle fingers on the arrow keys on the keyboard, indicating whether the orientation of the probe is the same (index finger; 'left arrow key') or different (middle finger; 'down arrow key') compared to the memory array grating. Feedback will be provided only in the cases of no response or an incorrect response, by presenting the word 'Wrong!' in red letters in the center of the screen for 1000 ms."

Regarding data trimming, we have included an additional section under *Analysis plan* (see *Data filtering* on p.21) describing the filtering process we will undergo in case we find evidence from speed-accuracy trade-off. Also, we mention that participants with an overall accuracy < 60% will be replaced and not be included either in the main analyses or during BF sample updating analyses. This section describes:

"*Data filtering. Participants with an overall accuracy close to chance levels (< 60% accuracy) in Experiments 1 and 2 will be excluded from analyses and replaced. The data of such participants will not be used during Bayesian sample updating nor for our main analyses. Additionally, we will exclude and replace participants in the case of technical or other difficulties, if data loss is greater than 20% of the total experimental trials. Further, the slowest and fastest responses will be removed from the analyses. To do so, we will filter each participant's responses and exclude any data that concern response times which are further than 3 standard deviations away from each participant's mean reaction time. Assuming that the reaction times of each participant are normally distributed, we expect less than 0.5% of the data of each participant to be excluded from the main analyses.*"

4) The analyses include a series of T-tests, which seems inefficient and "statistically costly" in terms of number of comparisons. Why are the authors not first resorting to repeated measures ANOVAs? The pattern of main and/or interaction effects would then be able to guide subsequent tests (with lower multiple comparison costs). The T-tests as currently presented are perhaps meant as planned comparisons -- if so, then the temporal relations between TMS timepoints are fully ignored (e.g., comparing early with late TMS timepoints).

Thank you for expressing these concerns, which we also share. Indeed, the reported t-tests are presented as the planned comparisons which are driven by the main research question (is the sensory visual cortex necessary during VSTM maintenance). In the current Stage 1 of the registered report, our planned comparisons are limited and specific to the statistical tests that focus on the reported hypotheses, which answer our main research question. We do in fact intend to perform repeated measures ANOVA and report them in an *Exploratory Analysis* section in during Stage 2 of the proposed registered report. Further, we are also interested in the differences between TMS timepoints, and given previous TMS studies (for a review see our pre-print Phylactou et al., 2021), we do anticipate a difference between these effects, where TMS effects induced earlier during the retention period will be stronger than TMS induced later in the retention period. However, this is a secondary question to our main research question. Given so, these analyses will also be reported in our *Exploratory Analyses.* In summary, given that this is a Stage 1 registered report, we only present the statistical tests which we will use to infer whether (or not) the sensory visual cortex is involved in VSTM maintenance.

5) The authors specify that participants will undergo a practise session of 24 trials. Given their status as "practise trials", it seems these data will not be analysed. However, I think a TMS-free condition prior to the main experiment would be an excellent way to assess baseline task performance, especially for

the non-stimulated hemisphere. Therefore, I would recommend that the authors include a (post-practise) baseline session free of TMS pulses (but perhaps after mounting the coil to the head).

Thank you very much for raising the issue of a baseline session, which is a concern that we have considered ourselves. After carefully considering our options, we think that including a baseline condition might not be necessary given the proposed control conditions in each of the experiments. In each of our two experiments, we provide two control conditions, and our proposed outcome neutral condition (Experiment 1) and sham TMS (Experiment 2), can address similar concerns to measuring baseline performance. Although a baseline session might indicate an anticipated performance without the introduction of TMS, we think this might be redundant since behavioral performance will be individually adjusted using staircase procedures, to approximate the individual threshold of orientation differences for 75% discrimination accuracy. Further, during practice, we will ensure that VSTM performance is equal to or greater than 75% accuracy as we describe on p.17, last paragraph starting on p.16:

"Furthermore, before the two main experiments, participants will carry out a practice block of 24 trials without TMS stimulation to familiarize themselves with the experimental procedure. If accuracy in the practice block is less than 75%, the practice block will be repeated until the participant reaches at least 75% accuracy. Participants will be replaced if after four practice blocks their accuracy remains below 75%."

As such, we do not expect the baseline session to be greatly informative since individual performance will be adjusted to an expected overall accuracy (75%). Thus, our outcome neutral condition (Experiment 1) and the sham TMS condition (Experiment 2), might provide a more appropriate indication for our main research question. Further, adding additional trials in our Experiments, could potentially lead to fatigue and/or practice effects, thus affecting the reliability of our results.

6) The phosphene procedure could perhaps be made more instrumental to the design and results. Phosphenes are arguably most easily elicited in pheripheral vision, and its visual field location of induced phosphenes could provide information about the specificity of TMS manipulation. The authors present their stimuli centrally, for good reasons, but I wonder whether final TMS target location and phosphene visual field location could be informative in explaining (lack of) effects. That is, would more centrally induced phosphenes impair memory maintenance more than peripheral phosphenes? If so, this would provide strong support for a topographically organized neural locus of the TMS effect (if any).

Thank you very much for pointing this out, allowing us to improve the proposed methodology. We agree with the statement that there is possibly a "*topographically organized neural locus of the TMS effect",* which is why phosphene percepts should overlap with the position where the stimuli will be presented (in our case, centrally). Following these suggestions, we will aim to induce the phosphenes as centrally to the visual field as possible, as we now report on p.15, first paragraph under *Procedure*:

"The TMS coil will be stabilized at the position where participants report phosphenes as close to the center of the visual field as possible, thus overlapping with stimulus presentation."

The topographical organization of the sensory visual cortex has been of great interest in the field, however we believe it is beyond the scope of the current proposed work. As such, we believe that for our line of work, in order to appropriately address this concern, we should focus on overlapping phosphenes with stimuli presentation, without attempting to explore the retinotopic nature of the sensory visual cortex.

7) Related to phosphenes: I would recommend that the authors check TMS and phosphene location between (one or more) trial blocks after experiment commencement in order to catch and correct drifts in location or (unexpected?) changes in visual cortical sensitivity to TMS.

Thank you for raising this concern, and we agree that by addressing this we improve the robustness of the proposed experiments. By following this suggestion, we now report that phosphene localization will be repeated half-way through each experiment (p.15, first paragraph under *Procedure*):

"Halfway through the experiments (after 3 blocks in Experiment 1, and after 4 blocks in Experiment 2), participants will be blindfolded again, and three single pulses will be induced on the mark, to confirm the induction of phosphenes and consequently stable coil placement. During this process, and if necessary, phosphene localization will be repeated, in order to adjust for possible drifts."

This will allow us to adjust any possible drifts and will also enable us to perform exploratory analyses if necessary. Such exploratory analyses might concern differences between the first and second half of the experiment, which might provide evidence for differences due to cortical sensitivity changes.

8) The authors do not mention analysis of response times (see also previous comment about response times). If the authors will not analyse response times, please explain why. If they will analyse them, analysis procedurs must be explained (including any trimming or filtering).

Similarly to your comment 4) in the *concerning methodology* section, reaction times will be analysed and reported in Stage 2 as *Exploratory Analyses*. In the current Stage 1 submission, we decided to limit our analysis plan to the most relevant tests regarding our main theoretical question of whether the sensory visual cortex is involved during VSTM maintenance. Given that *d'* is the most sensitive measure in our proposed experiments to measure VSTM performance, we focused our analysis plan on the main t-tests that allow us to infer the sensory visual cortex involvement in VSTM. Further, following similar suggestions by reviewer 1 (Dr Evie Vergauwe), we describe that we will replace participants performing close to chance (< 60% accuracy), and filter out the slowest and fastest responses. The data filtering process is described in detail in the revised registered report in the *Data Filtering* section on p.21, which we also transcribe below:

"*Data filtering.* Participants with an overall accuracy close to chance levels (< 60% accuracy) in Experiments 1 and 2 will be excluded from analyses and replaced. The data of such participants will not be used during Bayesian sample updating nor for our main analyses. Additionally, we will exclude and replace participants in the case of technical or other difficulties, if data loss is greater than 20% of the total experimental trials. Further, the slowest and fastest responses will be removed from the analyses. To do so, we will filter each participant's responses and exclude any data that concern response times which are further than 3 standard deviations away from each participant's mean reaction time. Assuming that the reaction times of each participant are normally distributed, we expect less than 0.5% of the data of each participant to be excluded from the main analyses."

9) Task stimuli: Some more details about the stimuli are required. What are the spatial frequency and contrast of the stimuli? Statement "(i.e. Gabor patch)" is not informative: Not all gratings are Gabor patches.

Thank you very much for pointing this out. In order to be clear about the details of the gratings, we have added the corresponding RGB colors that will be used to generate the Gabor patches, along with details regarding the frequency and distribution of the envelope used. On p.12, first paragraph under *Apparatus and Stimuli* (starting on p.11), we have added the following details:

"Stimuli will consist of either a red (RGB: 255, 0, 0) or a blue (RGB: 0, 0, 255) Gabor patch which will be oriented either horizontally or with a clockwise or counter-clockwise tilt from the horizontal axis, presented on a black (RGB: 0, 0, 0) background (Figure 2). The Gabor patch will consist of a gaussian envelope with a standard deviation of 0.39° (in degrees of visual angle), 0.001° frequency, and have a 1° diameter."

I hope that the authors can use my comments to further stregnthen their interesting manuscript and study design.

We would like to express our appreciation to the recommender and both reviewers for the valuable feedback. We are confident that the fruitful comments strengthen our study and the current registered report.