Dear Prof. Dienes,

We, once more, want to thank you and the reviewers for the constructive and insightful feedback. We are pleased to have received such a positive appraisal for the second stage of our registered report.

We have addressed the minor comments raised, as we present below. Your comments are presented in **bold font**, followed by our response (not bold).

Minor Revision

The reviewers are both very happy with your Stage 2, with some minor points to think about. Indeed, I agree it is well written and makes a clear contribution. Concerning your discussion of the point that a stopping rule to reach e.g. a BF of 3 (1/3) may lead to lack of evidence for other exploratory contrasts, or a lack of robustness, one solution is to stop at a higher threshold than one makes decisions for; e.g. the stopping rule refers to 6 or 1/6, but decisions are based on 3 or 1/3. That simple modification allows such decisions to be robust to scientifically reasonable changes to the scale factor.

We are contented knowing that yourself and the reviewers are happy with our Stage 2. We agree with your suggestion that the stopping rule could differ from the decision threshold. We have included this suggestion in the discussion of our paper, in pg. 42 (last paragraph from pg. 41), as follows:

"[…] Alternatively, future work could set two different BF thresholds for the stopping rule and the decision rule[2], to allow for additional sensitivity in case exploratory analyses are anticipated. For example, the decision rule to support a hypothesis might be set to a specific BF threshold (e.g., $1/3 < BF_{10} < 3$), but the stopping rule is set at a higher threshold (e.g., $1/6 < BF_{10} < 6$)."

Reviews

Reviewed by Evie Vergauwe, 08 Feb 2023 09:57

It was very exciting to see the results of this study, and overall, I think the authors did a great job! They nicely followed the preregistered plan, and I find their conclusions related to the preregistered analyses justified. There are a few issues that need to be considered, before recommending this manuscript, in my opinion:

We thank the reviewer for the positive comments.

1) Some small changes were made to the introduction, whereby additional references were added. I do not see a particular problem with it, but I am not sure to what extent changes like these can be made to the introduction at this point in the process.

We are thankful that this was pointed out. We removed the additional references that were added, since, according to PCI guidelines (as published in Section 2.10 in PCI guide for

authors; https://rr.peercommunityin.org/help/guide_for_authors), such changes are to be avoided.

2) One of the footnotes mentions "The number of trials in Experiments 1 and 2 differ from the registered trial numbers, due to correcting a mistake in the calculation of the required trials (360 instead of 432 in Experiment 1 and 512 instead of 572 in Experiment 2). This error was corrected prior to any data collection and after receiving the recommender's approval on 22 June 2022". I think it would be good to specify explicitly that the modification is, in fact, an *increase* in the number of trials (and thus more data). Also, for Experiment 2, the main text mentions 576 trials, but the footnote mentions 572 trials – this should be the same number.

We appreciate the reviewer's suggestion of rephrasing our footnote (see *Notes* section in p. 52) and for identifying the typo with the trial numbers. We have now updated the footnote to include the suggestion, and also correct the typo (572 instead of 576 trials), and we also provide this update below:

"The number of trials in Experiments 1 and 2 have been increased compared to the registered trial numbers, due to correcting a mistake in the calculation of the required trials (432 instead of 360 that were registered in Experiment 1 and 576 instead of 512 that were registered in Experiment 2). This error was corrected prior to any data collection and after receiving the recommender's approval on 22 June 2022."

3) On p. 21, the following sentence is hard to understand (does the second part need to read "we would recruit" instead?, or should the "if" be a "because"?): "If any of the three BFs did not reach the stopping rule of > 3 or < 1/3, we recruited four more participants and repeated the analyses"

We thank the reviewer for bringing this to our attention. To make the sentence easier to understand, we rephrased it as follows (pg. 21, second paragraph in *Sampling Plan* section):

"[…] If any of the three BFs did not reach the stopping rule of > 3 or < 1/3, we continued with data collection, as follows: we recruited four additional participants and performed the analyses again. [...]"

4) Figure 3 and 5: The names of the different conditions are put as the titles of the Y axis, which I find confusing. The title of the Y axis should be the measure that is displayed (here: Bayes factor). And then the names of the conditions can go above or below the graphs.

Figures 3 and 4 have now been adjusted according to the reviewer's suggestion, who we thank for proposing this adjustment. Here, we wish to clarify that we assume that the reviewer was referring to Figures 3 and 4 (not 3 and 5), as these two figures are the only two that have the Bayes Factor in the Y axis.

5) There are some model comparisons and interpretations of results that deserve some attention (these concern only exploratory analyses):

5.1: p. 33 "The Bayesian rmANOVA generated the highest BF10 for the model that included only the TMS site factor (BF10 = 3.46), showing that the observed data are better represented by considering the ipsilateral and contralateral differences)." –> the fact that the best model of the data included only the TMS site factor, shows that the data are best explained by *only* considering the differences between the ipsilateral and contralateral sites.

5.2: p. 33: the authors conclude that the Timing factor is not adequate to explain the observed data, because the Timing-only model was much worse than the null model (Experiment 1). However, that may not be the most appropriate model comparison to assess the evidence in the data for a main effect of Timing factor. Given that, for the main effect of Site, the authors start by stating the best model of the data (i.e., Site-only), it makes more sense to me to then examine how much worse this model fares if one now adds the Timing factor. Based on the values in Table 4, this would show that the data are inconclusive when it comes to the Timing factor: 3.46/3.02 = 1.15. There is a similar issue for Experiment 2, where the Timing-only model was compared to the null model, to examine the main effect of Timing, rather than by adding the Timing factor to the best model. Adding the Timing factor to the best model gives some evidence against a main effect of Timing (9.06/1.79=5.06) in Experiment 2. Related to these points, I think that some statements in the discussion may be too strong when it comes to the absence of a Timing main effect. Furthermore, it seems that some of the statements made in the general discussion about Timing effects are actually about the *interaction with Timing*, rather than the main effect of Timing. If that is indeed the case, the BF's of the relevant model comparisons should also be reported in the results section.

We agree with the suggestion of the reviewer that to justify our claims regarding the timing factor, it is necessary to further explore how including this factor affects the evidence for our model. For this reason, we have now included additional exploratory results, describing the analysis of the effects, which are derived from our Bayesian rmANOVA. In detail, we now report, for each factor of the model a BF_{incl}, which describes the calculated likelihood ratio of the change from prior odds to posterior odds for each factor in the model averaged by all the models that include each factor. Given that for both Experiments 1 and 2, the Timing factor, and any interaction including the Timing factor points towards a null effect (all Timing $BF_{incl} > .73$), we are confident that parallel to the reviewer's suggestion, the statements that we discuss regarding the Timing effects are now supported by the additional results that we report.

These results are reported in pg. 33, (last paragraph starting in page 32) (Table 5 in pg. 60), for Experiment 1 as below:

"[…] To further explore the model that better represents the data, we conducted analysis on the factor effects by calculating the likelihood ratio representing the change from prior odds to posterior odds for each factor in the model averaged by all the models that include each factor (BFincl). The BFincl for all factors and interactions are provided in Table 5. In detail, the inclusion of the TMS site factor resulted in the highest BFincl (BFincl = 23.01), but including the TMS timing factor in the model, resulted in lower posterior odds, pointing to

evidence against (i.e., BFincl < 1) a model including the TMS timing factor (BFincl = .73) or an interaction of TMS site and TMS timing (BFincl = .34). […]"

And in pg. 34, (last paragraph starting page 33) (Table 7 in pg. 62), for Experiment 2 as follows:

"[…] As with Experiment 1, we performed an analysis of effects by calculating a BFincl for each factor and interaction included in the model. The BFincl resulting from this analysis are presented in Table 7. Specifically, the highest BFincl was produced by the TMS condition model (BFincl = 31.45), followed by that of the TMS site model (BFincl = 15.45). Models including solely TMS timing, or TMS timing interactions resulted in low BFincl (all BFincl < .36; see Table 7 for details). […]"

Reviewed by Vincent van de Ven, 16 Feb 2023 11:46

I reviewed the latest version of a Stage 2 manuscript of the authors' study, which comprises 2 experiments that utilize double-pulse TMS to modulate visual cortical activity during the maintenance phase of a change detection task. The methodology and planned analyses were already approved at Stage 1, in which I was also involved in one of the reviewing rounds.

At this moment in the manuscript development, I can be short in stating that I find the manuscript a very impressive and enjoyable read. The Introduction is authoritative and well balanced, as the authors carefully describe the background literature in a detailed and disciplined manner, with emphasis not only on previous TMS work but also on the recent debate about whether the visual cortex supports (short-term) memory formation / consolidation. Methods are minutely described in relevant detail and experimental design decisions are well motivated. The Results are presented in a thorough, rigorous but also easily readable fashion, where I find the empirical evidence strong and convincing. The two experiments provide a strong combination of evidence, by replicating main results as well as extending the methodology to sham stimulation, which is an important but (in my view) also somewhat controversial procedure in TMS research. The Figures are clear to me (although lettering is sometimes a bit hard to read in the provided PDF). The Discussion carefully considers the findings in light of previous research and points to relevant future steps to gain further insight in how we retain visual information in memory. Overall, the manuscript is long but reads very easily. In this sense, I find the manuscript a strong and important contribution to the current literature and a stepping stone to future implementations.

We express our appreciation for the positive feedback of our work by the reviewer.

The only two comments I would like to make are:

1) The results for 1000 msec timepoint of stimulation is less strong (BF just above 3) in comparison to other timepoints. While this surely can be considered as evidence for a "late" consolidation window in visual cortex (as the authors seem to do), I would have liked to see a bit more consideration of this effect. Especially in light of some studies also

showing smaller (or null) effects for late timepoints (in TMS as well as in visual memory masking). Perhaps, at a BF just above 3, the glass is proverbially half full or half empty, depending on one's preference in this matter.

We agree with the reviewer that with the BF just above 3, it can be argued that evidence is debatable. This can also be partly attributed to our sample updating design, where the stopping rule was set to BF = 3. For example, if a higher threshold was set, then the additional data could have provided a clearer picture. This is acknowledged in our limitations, which following the recommender's suggestion has been adjusted in this revised version in pg.41, last paragraph, as follows:

"Lastly, despite the benefits of the sample updating with stopping rule design (Rouder, 2014), this approach can also be subject to limitations. In our study, the stopping rule was focused on our registered analyses, which enabled us to tailor our sample size accordingly, so that we gathered adequate evidence regarding our hypotheses (see Table 1), as reflected by the BF, while preserving resources (Rouder, 2014; Schönbrodt et al., 2017; Schönbrodt & Wagenmakers, 2018; van Ravenzwaaij & Etz, 2021). However, the sole focus on the registered analyses and their proposed prior distributions (see Table 1), resulted in some inconclusive results in the exploratory and robustness analyses, since the predefined threshold was not reached ($1/3 < BF10 < 3$). Even though these exploratory analyses were not the focus to drive the theory of the two experiments presented here, it is possible that with additional participants, the proposed BF10 threshold for additional analyses would have been reached (e.g., see Figure 3C). As such, future studies using Bayesian designs could rely on different approaches for sample size determination, such as simulations that can inform the minimum required number of participants for various study designs (Fu et al., 2021; Phylactou & Konstantinou, 2022; Schönbrodt & Wagenmakers, 2018). Alternatively, future work could set two different BF thresholds for the stopping rule and the decision rule[2], to allow for additional sensitivity in case exploratory analyses are anticipated. For example, the decision rule to support a hypothesis might be set to a specific BF threshold (e.g., $1/3 < BF_{10} < 3$), but the stopping rule is set at a higher threshold (e.g., $1/6 < BF_{10} < 6$)."

In addition, given the failure to find any evidence favoring a TMS timing effect (or interaction) has led us to the conclusion that even though the 1000 ms seems smaller in terms of BF, it is not, at least in our experiments, distinguishable from earlier TMS effects. In this revised version of our Stage 2 report, we provide complimentary analyses to our ANOVA supporting the absence of a Timing effect, following the suggestion of a different reviewer (Evie Vergauwe).

These additional analyses are reported in pg. 33, (last paragraph starting in page 32) (Table 5 in pg. 60), for Experiment 1, and are also provided below:

"[…] To further explore the model which better represents the data, we conducted analysis on the factor effects by calculating the likelihood ratio representing the change from prior odds to posterior odds for each factor in the model averaged by all the models that include each factor (BFincl). The BFincl for all factors and interactions are provided in Table 5. In detail, the inclusion of the TMS site factor resulted in the highest BFincl (BFincl = 23.01), but including the TMS timing factor in the model, resulted in lower posterior odds, pointing to

evidence against (i.e., BFincl < 1) a model including the TMS timing factor (BFincl = .73) or an interaction of TMS site and TMS timing (BFincl = .34)."

And in pg. 34, (last paragraph starting page 33) (Table 7 in pg. 62), for Experiment 2 as below:

"[…] As with Experiment 1, we performed an analysis of effects by calculating a BFincl for each factor and interaction included in the model. The BFincl resulting from this analysis are presented in Table 7. Specifically, the highest BFincl was produced by the TMS condition model (BFincl = 31.45), followed by that of the TMS site model (BFincl = 15.45). Models including solely TMS timing, or TMS timing interactions resulted in low BFincl (all BFincl < .36; see Table 7 for details). […]"

2) In the Discussion, I would have liked to see a reconsideration (however brief) of the current debate about the role of visual cortex in short-term consolidation / retention in light of the current findings. In her reviewing work, Xu (and colleagues) includes previous TMS findings in her considerations that visual cortex is not suited to store visual memories. One could have perhaps linked back to her argumentation and consider in how far this view is modulated by the current evidence.

We are thankful to the reviewer for the suggestion of linking back to the review against sensory recruitment by Xu (2017). We agree with the reviewer, and thus, in the revised Stage 2 we have adjusted, in pg.39 (last paragraph starting pg. 38) as we transcribe below:

"[…] Notably, a recent review (Xu, 2017) argued that the stronger effects for earlier TMS found in some previous studies (Rademaker et al., 2017; van Lamsweerde et al., 2017) can be taken as evidence against the storage of information by the sensory visual cortex during VSTM. This argument was further complemented by the null finding in the study of van de Ven et al. (2012) for TMS at 400 ms. However, we argue that a weaker effect during later stimulation does not correspond to the absence of an effect. Contrary, as reflected by our results, even though the likelihood of the evidence is lower for later stimulation, the effects of TMS cannot be differentiated based on timing of the stimulation (see also Phylactou et al., 2022). Along these lines and in contrast to the argument of Xu (2017), we propose that, taken together, evidence from TMS supports the idea that sensory visual cortex is an essential part of the network involved in VSTM."

However, I offer these comments are optional for the authors to consider. I do not wish to hold back this well written and elaborately described manuscript on just these points.

We once again thank the reviewer for their positive review, however, we found the comments valuable for the improvement of our Stage 2 report and did our best to address them.