**Causal temporal dynamics of task-relevant rule and stimulus selections**

*Jade Buse Jackson Runhao Lu Alex Woolgar*

Stage 1 Review for PCI-RR - ArticleID #570

**Dear recommender,**

**I have carefully read the Stage 1 report by Jackson and colleagues, who propose an investigation, with TMS, of the role of two frontal brain regions at two different timepoints of task relevant rule and stimulus selections. Their proposed study has the potential to provide causal evidence for the role of dlPFC and/or dmPFC during task processing.**

**I applaud the authors for their efforts, as the current report seems thorough and well thought of. Below, I provide some suggestions, which I think will help strengthen the current registration and subsequently benefit the conduct of the study.**

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

The authors give sufficient background to support the validity of their research questions, which are also strengthened by their (unpublished) MEG data presented in the report.

One area I find lacking, and that could benefit from additional detail, is the rationale of the different stimulation timepoints, especially the ‘late’ stimulation. In detail, the authors mention “*[b]ased on previous literature and our MEG data, we also anticipate that at the earlier stimulation timepoint and in the active dlPFC condition compared to sham dlPFC we will observe a higher percentage of rule-based errors, while at the later stimulation timepoint we anticipate a higher percentage of stimulus-based errors*” (p.4), however not much information is provided about these expected temporal differences. Any previous findings that can support this would help make the authors’ case stronger. I find this important, especially since the MEG data (as the authors also acknowledge in their report) do not strongly support temporal differences in their decoding accuracy analyses for prioritization of relevant and irrelevant information.

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

The authors seem to have given a lot of thought for the 36 proposed hypotheses. For the purposes of a Stage 1 RR, I find that 36 hypotheses complicate the design of this report. I acknowledge the fact that the authors made an attempt to register each, but the report will benefit if it remains focused on the main research driven questions of the study. The issue with the large number of hypotheses also becomes even more complicated given the potential results and interpretations of the ANOVAs (I discuss this issue in detail in section 1C) and the multiple dependent variables (i.e., RT and ACC; I discuss this issue in detail in section 1D).

The authors are interested in (i) the causal role of the dmPFC and/or the dlPFC, (ii) the potential temporal differences in their involvement, and (iii) their causal involvement in rule and/or stimulus processing, (iv) temporal differences in rule/stimulus processing for dmPFC and dlPFC. The authors also have registered (v) analyses for testing for TMS artifacts. The hypotheses in (i) and (iii) seem to be supported by the theory and findings the authors have provided, however as I previously mentioned, the support for the temporal differences is not strong. That is not to say to drop (ii) and (iv). If the authors make a strong case for these hypotheses they may wish to keep/update them. However, in the current context, I would advise the authors to treat these hypotheses as exploratory. The TMS artifact hypotheses, in their current form could also be treated as exploratory, since the authors’ provided interpretations are mainly related to the TMS questionnaire per se, and not their theoretical driven research questions. Alternatively (and what I would advise), the authors can update their interpretations so that Q5 can be considered as a quality check analysis. The authors do indeed mention *“[…] in a way that mirrors the results under Q1 then this would weaken our overall interpretation of the results*” (p.26), which points towards a quality check analysis, but “*weaken our overall interpretation*” is vague. The authors should consider providing interpretations in terms of what would validate and invalidate their registered analyses.

Additionally, I agree with the authors that the most appropriate approach to explore these questions statistically would be with the use of ANOVAs. However, the authors could consider limiting their registered analyses to specific contrasts that can be answered by very specific contrasts (e.g., a specific *t*-test). For example, for the registered analysis for H1 and H2 this t-test could be the respective sham vs. active TMS, and so on. The ANOVAs could be treated as exploratory to explore potential interactions (e.g., the timepoint differences etc.). If the authors do wish to keep the ANOVAs as their registered analyses, then the provided interpretations will require further detail to reduce all research degrees of freedom (see my comments in the following sections).

Further, from what I understand, H2 and H4 test the same contrast but with a different direction. Why not test this with a single two tailed test instead of two one tailed tests? This will also result in reduced analytic flexibility.

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

Overall, the study is thoroughly and carefully designed. Some aspects of the methodology and analysis can be improved, and I offer my suggestions in the sub-sections below:

**Evidence Threshold:**

It is unclear what the threshold for accepting the evidence is. The authors mention that they will use updating with BF > 6, but it is not clear if BF > 6 is also the evidence threshold. It is possible to have different thresholds for the stopping rule and the evidence, and advisable if the authors plan to conduct exploratory analyses. On a similar note, the authors mention “*intermediate Bayes Factors*”, though it is not clear how “intermediate” is defined.

**Sample Size Justification:**

The authors propose updating with a stopping rule of BF > 6 for their sample size. This is tricky as the authors will rely on multiple ANOVAs for their stopping rule, which require further detail for clarification. For example, the ANOVA with the *Stimulation* condition and the *Timepoint* condition will result in multiple models with different probabilities P(model|data). As such, different BFs can be computed for different model comparisons. How is the stopping rule going to be implemented in this case? Does BF > 6 correspond to the comparison between the interaction model and the null model? What if the interaction model doesn’t reach the stopping rule but other models do? What if the interaction model reaches the stopping rule compared to the null but there is evidence against it in comparison with other models? This could be avoided if the registered analyses relied on simpler models (e.g., *t*-tests as suggested earlier). If the authors wish to rely on the ANOVAs for the stopping rule, then they could describe in detail which models they are comparing to generate the BF [e.g., Pstimulation\*timepoint(model|data) / Pnull(model|data)] and how they would interpret possible evidence of other model comparisons [e.g., Pstimulation\*timepoint(model|data) / Pstimulation (model|data) or Pstimulation\*timepoint(model|data) / Ptimepoint(model|data), etc.].

The authors describe situations where participants will be excluded from analyses (e.g., p.5). Can the authors clarify whether they will replace these participants or whether these participants will be accounted for in their sampling plan (e.g., will they be considered for the minimum and maximum *n* size)?

**Experimental procedure**

During the proposed task, a visual stimulus would be presented for 117 ms, followed by a 3200 ms response window. TMS will be applied as a 230 ms train at either 150 ms or 700 ms, which means that a TMS pulse can be applied as late as 930 ms. Will participants be allowed to make a response during this time? If so, this might be problematic, especially at the late stimulation condition, as participants might be able to respond before the TMS train has been applied. I am not sure whether a dedicated response screen (i.e., delayed-response task) can be used to resolve the issue, as some might argue that in such a case the task will be resembling a working memory task, which the authors might wish to avoid. In all honesty, I am unsure how this possible issue can reliably be resolved, but maybe the RT data from the MEG study could provide insight as to whether this might be an issue.

In addition to the TMS artifacts that will be measured, I would suggest also recording whether participants distinguished between active and sham TMS. I understand that the authors might not want to explicitly ask this at the end of every session as with the questions in Appendix, though they could do this at the end of the experiment, and test for differences between the group that noticed the difference and those who did not.

**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 authors describe their methods and analyses in great detail. Further suggestions are provided below to reduce flexibility and increase replicability.

**Analytic Flexibility**

As mentioned above, some analyses seem repetitive (e.g., H2 and H4) and can be tested by a single test. Providing a single hypothesis, with its potential interpretations and a single test, will reduce analytic flexibility. For example, in H2 the authors hypothesize a greater role of dlPFC but in H4 a greater role of dmPFC. These hypotheses contradict each other. The authors should provide their expectation and test this relationship with a single test.

Another analytic flexibility issue relates to the registration of multiple dependent variables. For example, the authors plan to analyze both ACC and RT to draw their conclusions. Even though it is understandable to test for effects on both variables, in the context of an RR, this raises potential flexibility and interpretation issues. For example, is evidence for one (RT **or** ACC) adequate to draw conclusions or is evidence for both (RT **and** ACC) required? How will results be interpreted if the findings are contradictory (e.g., faster RT but lower ACC, or slower RT but higher ACC)? Does one variable have bigger weight than the other? The authors could consider (i) relying on one variable for registration and treating the other as exploratory, or (ii) pulling ACC and RT together to a single variable, similar to a speed-accuracy trade-off approach (e.g., Liesefeld et al., 2019; <https://doi.org/10.3758/s13428-018-1076-x>)

**Replicability**

The authors should provide a data availability statement.

The report will benefit from additional details regarding the stimuli, such as the stimulus size, cue size, and the viewing distance.

**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).**

Overall, the authors present a thorough and well-thought of design that ensures that the obtained results will test the planned hypotheses. The methods include a TMS questionnaire to capture potential TMS artefacts, which could serve as a sufficient quality check for the study.

I hope that the authors find my suggestions insightful.

Respectfully,

Phivos Phylactou