Dear Professor Chambers,

Thank you for your supportive comments on our first submission and thank you to yourself, Yuki Yamada, and the reviewer, Phivos Phylactou, for the detailed and constructive reviews. We have carefully considered your recommendations and revised our study protocol accordingly as described in detail below.

In addition to making the requested changes, we have decided to include simultaneous EEG recording in the TMS sessions as this will provide complimentary neural information, and a richer dataset, for what will be a labour-intensive experiment. We now specify in the registered report that participants will be setup with simultaneous EEG. However, we do not plan to register any EEG analyses as we will keep this dataset separate and analyse it at a later stage. The addition of EEG necessitated two changes to the protocol 1) we now will acquire motor thresholds manually rather than with the grid method due to practical constraints and 2) participants will now alternate between two stimulus response-mappings during the main task allowing us to dissociate the stimulus decision and the response in the neural data. These changes are included in the report.

Due to funding constraints, we have had to start the hiring procedure for the research assistant who will carry out data collection, with a planned start date in June. We appreciate your offer to consider this revision yourself, given your familiarity with the topic, without sending it back out for specialist review to expedite the timeframe (if you feel we have addressed the points thoroughly). This would be extremely helpful in ensuring that we can hold off data collection until after IPA.

Your sincerely,

Jade Jackson, Runhao Lu, and Alex Woolgar

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.

Comparing the stimulation timepoints is one of our key interests. We now reference previous work from our lab (Goddard, Carlson, & Woolgar, 2019) which used a similar task design and demonstrated earlier peak frontal decoding of the relevant rule compared to the relevant stimulus features. We now also reference an MEG study (Quentin et al., 2019) which implemented a visual working memory task and showed that after the cue (rule) display there was initial peak decoding of the rule and a period of maintenance followed by decoding of the relevant stimulus features ~600ms later (in the section MEG data: Study design, see added text below).

"Previous studies have indicated that rule and relevant stimulus processing have distinct temporal profiles (Goddard et al., 2019; Quentin et al., 2019). For example, in Quentin and colleagues' (2019) study, after a rule cue, there was immediate coding of the rule which was maintained throughout the trial, followed by coding of the relevant stimulus features starting from ~600ms after the rule cue presentation. However, we chose to derive temporal predictions for our specific paradigm and used the exact same task in the MEG that we planned to use in the proposed TMS study."

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.

Thank you for these suggestions. We have reduced the number of registered tests which address our key research questions to fifteen Bayesian paired t-tests. We have kept the registered tests which investigate temporal differences as this is one of our key questions. We have changed one-tailed tests to two-tailed where appropriate, i.e., the comparison of active dmPFC to active dlPFC. We have also combined reaction time and accuracy into a single dependent variable in line with your comments as we do not have strong hypotheses for using one over the other as a dependent variable. Finally, we have further developed our quality check and have included, in our design table, a clear indication of the circumstances under which these data would invalidate our registered analyses.

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.

Our stopping threshold is BF > 6 for the alternative and BF < 1/6 for the null. However, for interpretation, for example with our data checks, if we reach our maximum n, or if we conduct exploratory analyses, we will interpret BFs between 1/3 and 3 as showing insufficient (intermediate) evidence, BFs < 1/3 or BFs > 3 as moderate evidence, and BFs < 1/10 or BFs >10 as strong evidence. This is now clarified in the section "analysis protocol" and in the study design table. When we refer to "intermediate Bayes Factors" we mean any value of 1/3 > BF < 3. This is now clarified in the study design table.

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., $P_{stimulation*timepoint}$(model|data) / P_{null}(model|data)] and how they would interpret possible evidence of other model comparisons [e.g., $P_{stimulation*timepoint}$(model|data) / $P_{stimulation}$ (model|data) or $P_{stimulation*timepoint}$(model|data) / $P_{timepoint}$(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)?

As specified above, we have modified our registered analyses from ANOVAs to planned Bayesian paired t-tests as these tests more directly address our primary research questions. We have also clarified in the section *Participants* that participants who are removed will be replaced (and not considered in the minimum/maximum *n*).

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.

The late train will start at 700ms, with the three pulses at 700, 777 and 854ms. We will remove trials where participants respond at any point before the last TMS pulse. We have plotted the individual subject RTs from the corresponding MEG data and estimated that we will need to remove approximately 11% of trials due to participants responding before the last TMS pulse. We have now specified in the study design table that we will remove trials where participants respond before the last TMS pulse.

We anticipate that the quality check questionnaire will capture any somatosensory differences between the sham and the active TMS conditions. However, it is possible that there may be a psychological impact if participants are aware that they are receiving sham rather than active stimulation. We recently acquired some pilot data with our sham coil where participants were asked to guess at the conclusion of the experiment, which conditions were placebo, and which were real TMS. While 50% of participants rated the sound-matched sham coil as a placebo condition, 64% of participants also rated the real TMS condition as placebo. This pattern is difficult to interpret, and we think it reflects the question being a poor indication of their experience, so we have decided not to include this question in this study.

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)

We have changed one-tailed tests where appropriate to a single two-tailed test. Further we appreciate the reviewer's commentary on accuracy and RT interpretation, and as we do not have a strong prediction concerning one or the other we have decided to combine them into a single measure (IES). We have now specified this in the section “proposed analyses” and in the study design table.

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.

We have now provided a data availability statement and have included details on the stimulus and cue size, and the viewing distance (section: stimuli).

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

Recommender comments

1. What happens if participants respond during or before the late TMS? Will TMS still be administered? How (if at all) will this be taken into account in the analysis? How was RT taken into account in the MEG analysis? It strikes me as a significant interpretative concern if the TMS is delivered during or after the response is executed, as it logically would be unable to influence cognition.

We agree with these comments and as specified above in response to the reviewer prior to analysis, we will remove any trials where participants responded before the last TMS pulse. TMS will still be administered on these trials.

2. Please clarify the timing of TMS pulse trains as there appears to be potential discrepancy between the details in Figure 2 (trains starting at 250ms or 800ms) and the description on p17 (trains starting at 150ms and 700ms).

The timings are 150ms and 700ms, we have corrected it in Figure 2.

3. It seems to me that if DLPFC or DMPFC TMS impairs attentional selection (or even perception) of the cue it could produce a rule error (or RT slowing) without affecting rule processing *per se*. Therefore I find myself wondering if the design would benefit from an additional negative control to confirm that prefrontal stimulation leaves perception/attention of the cue unaffected. I will leave you to consider how best to achieve this, but one possibility could be to insert some trial blocks in which participants need to discriminate the cue type as quickly as possible (e.g. & vs ! or $ vs %), and a Bayesian t-test could be used to search for any effect of active vs sham TMS on RT and error rates. I note that you do give participants the option to "press a fifth button with their left hand if they did not see the stimulus or the rule symbol", which would capture a very large effect of TMS on lower level processes, but any such disruption of attention/perception is likely to be too subtle to be detected using such a response choice. In general there is risk, as with all TMS studies, that because the cognitive task being used involves quite high-level processing, that at least some TMS-induced deficits observed on the task must be originating at a similarly high level, when it is possible that any lower level disruption could have knock-on effects. These potential lower-level causes need to identified and controlled as much as possible.

Thank you for this suggestion. We will now ask participants to complete a separate task at the end of each TMS condition (active dlPFC, sham dlPFC, active dmPFC, sham dmPFC). In this task and on each trial, they will be presented with one of the four symbols (&, !, $, or %, for 117ms) and will receive a train of three TMS pulses at 13 Hz commencing at 150ms. Following the train of pulses and at 310ms they will be presented with a response screen with the four symbols and will need to indicate as fast and as accurately possible which symbol they saw by pressing the corresponding button. We will combine accuracy and RT into a single measure and compare the active and respective sham conditions to each other as well as the active conditions to each other. If perception of the cue is impaired in a way that mirrors the main hypotheses e.g., worse performance in the active dlPFC compared to the sham dlPFC condition in the cue perception task and a higher percentage of rule errors in the active compared to the sham dlPFC condition in the main task, then we would not be able to conclude that TMS interfered with rule processing. We would instead infer that TMS disrupted (in the main task) either the perception/attentional selection of the cue, or a combination of this, and rule processing. We have included details of this control task in the section "structure of each section", "data checks" and in the data checks table.

4. Please fully specify the interpretative consequences in any differences between sham vs active TMS in the stimulation artefact analyses (H37). You note that it will weaken the interpretration of the results (which is a important starting point), but it is crucial to make clear by how much it will do so. In other words: which outcomes of this analysis (if any) would render the results of the main hypotheses completely inconclusive? Without a clear and precise interpretative plan, I fear it will be highly tempting to dismiss any artefact differences. Knowing how much work goes into such large-scale TMS studies, I know I would certainly be tempted to do so myself!

We agree with this comment and the similar suggestion from the reviewer; accordingly, we now have a clear quality check plan for the questionnaire data. We will employ Bayesian paired t-tests to compare the active TMS conditions to their respective sham conditions, and the two active TMS conditions to each other, in each of the somatosensory dimensions. If there is moderate evidence for any one of these t-tests (BF > 3) we will then test if there is evidence that a change in these scores (e.g., a change in the experience of pain) predicts the change in performance in the main task between the corresponding TMS conditions. If there is evidence towards this, then we will consider the intercept of the regression line: if there is evidence that the intercept is different from zero, this will indicate a difference that remains even after accounting for the difference in subjective experience. However, if there is evidence for the null (the intercept is not different from zero) then the conclusions of the main test will be invalidated. We detail this in the section "data checks" and in the data checks table.

5. Will participants wear hearing protection (e.g. ear plugs)?

Yes, they will, this is now specified in the report.

6. Have you done any piloting to explore risk of blink artefacts due to facial nerve stimulation? In our own studies we sometimes found that some participants can be susceptible to these artefacts, and unfortunately timed blink artefacts could produce behavioural results that look like those produced by cognitive interference (particularly for the early TMS epoch). If you have eye tracking available, this would be ideal use-case to detect and exclude any trials in which blinks occured during the cue/stimulus presentation. At a minimum, it may be a good idea to check in session 2 that the active TMS doesn't cause blinks in each participant.

We agree that it is possible that stimulation will cause some participants to blink depending on where the coil is situated on their head. However, the early TMS train commences 150ms after stimulus onset and stimulus duration is 117ms so any blinks will follow stimulus presentation. As the first TMS pulse will be delivered after stimulus presentation, and we are already using the questionnaire as a quality check to look for differences in somatosensory experience between the TMS conditions, we have decided not to include eye tracking.

7. A general comment: but given the complexity of the design, please pass through everything and check that the exclusion (and participant replacement) criteria are as comprehensive as possible, as these are generally not possible to change for confirmatory analyses after Stage 1 in-principle acceptance.

We have done so.

Managing Board review (provided by Yuki Yamada)
The methods are very detailed, technical and skillful information is provided, and I could not detect any major problems here. However, I felt that there are too many hypotheses. In confirmatory research, hypotheses for testing need to be theoretically justified and validated, but I doubt that all 40+ hypotheses here have such a background. I rather got the impression that this study is exploratory in nature. It would be good if the authors could clarify which (style) of research, exploratory or confirmatory, this study is. Regarding the sample size, I could not find any clear rationale that the minimum sample size should be 24. Also, there is a discrepancy between the Participants section (N=60) and the Proposed analyses section (N=56) regarding the maximum sample size.

In line with both the reviewer and managing board's comments we have reduced the number of registered hypotheses (which address our main research questions) to 15 Bayesian paired t-tests. This study is confirmatory rather than exploratory as we have provided justification for our hypotheses in the introduction. Also, we thank you for pointing out this error, the maximum sample size is n=56 as we will collect data in sets of 8 after each full counterbalance of our TMS conditions. The minimum sample size is a multiple of our counterbalance number, however, false positives are higher at a smaller n, so instead of starting with a minimum of 8 or 16 we started with a minimum of 24. We have included this justification in the main report.

References

Goddard, E., Carlson, T. A., & Woolgar, A. (2019). Spatial and feature-selective attention have distinct effects on population-level tuning. *bioRxiv*, 530352. doi:10.1101/530352

Quentin, R., King, J.-R., Sallard, E., Fishman, N., Thompson, R., Buch, E. R., & Cohen, L. G. (2019). Differential brain mechanisms of selection and maintenance of information during working memory. *Journal of Neuroscience, 39*(19), 3728-3740.