ROUND 1

Chris Chambers

Three reviewers have now completed an initial evaluation of your manuscript, and I have also read it with interest myself. Overall, the reviews are encouraging about the potential for Stage 1 acceptance, following a thorough revision to strengthen various elements of the study design and presentation. Among the various comments, the reviewers highlight the need for clarifications to the study rationale, procedural details, and analysis plans. Two of the reviewers suggest adopting an alternative (or at least complementary) analysis plan involving Bayes factors, and I would very much encourage you to consider this because the study outcomes will then be more informative, regardless of the results. If you eventually adopt both frequentist and Bayesian inferences, be sure to specify which outcomes (the Bayesian or frequentist) will shape the conclusions. Other comments should be straightforward to address by adding minor details to the manuscript or noting in your response where a particular detail was missed (e.g. I note that point 4 of the 2nd anonymous reviewer -- definiition of S1 -- is already stated on p7).

I look forward to receiving your revision and response, which I will return to the reviewers for re-evaluation.

R: Thank you very much for your encouraging and thoughtful comments. We provide a point-by-point response to the Reviewers' comments below, with changes in the manuscript highlighted in bold. Regarding the suggested Bayesian approach, we have now included it as complementary to the statistical analyses we proposed. We believe that the revisions have greatly improved the rigor of our manuscript, which we hope is now suitable for Stage 1 acceptance.

Looking forward to hearing from you at your earliest convenience.

Best Regards,

Agnese Zazio (on behalf of all authors)

REVIEWER 1

Review of Stage 1 registered report by Zazio et al. The points below are organised according to the stage 1 criteria.

R: Thank you for your important comments, which helped us in improving the quality of the manuscript.

1A. Validity of the research question. The authors postulate differences in associative learning mechanisms in borderline personality disorder. It's not clear why they hypothesise that these differences should only be present for social associations, i.e. tactile mirroring. A more general deficit in associative learning should have widespread effects on cognitive functioning. Please

present evidence for such a deficit or a clearer rationale for why BPD patients should have specific differences in tactile mirroring.

R: Thank you for giving us the chance to clarify our rationale. Our hypothesis of associative learning dysfunction within TaMS in BPD arises from the evidence showing alterations in mirror systems activity (Mier et al., 2013; Sosic-Vasic et al., 2019) and from the observation that the development of mentalization and empathic abilities appears to rely on early associative learning (Grosjean & Tsai, 2007), which may explain the difference in empathic levels compared to controls (Harari, Shamay-Tsoory, Ravid, & Levkovitz, 2010; Martin, Flasbeck, Brown, & Brüne, 2017). Specifically, we do not hypothesize a deficit in associative learning per se. Here, we aim at detecting whether it is the case that, within a specific system, i.e., the tactile mirror system (TaMS), associative learning is altered in BPD. Clearly, the experimental design we propose at the present stage will not be able to establish whether the possible alterations we observe are specific for the TaMS or whether they extend to other cognitive domains, as the PAS protocol we propose is cross-modal and relies on the TaMS. In case of positive findings, it will be important to address this aspect in future studies. We have now clarified this point at p. 4.

1B. Proposed hypotheses. The authors propose that cm-PAS will improve tactile acuity but 'decrease performance' on the visual-tactile spatial congruity task. Please specify more clearly how the decrease in performance will be indexed, i.e. as an increase in response times on incongruent trials, a decrease in response times on congruent trials, or both?

R: Thank you for this comment. Indeed, a "decreased performance", i.e., greater $\Delta RT_{incong-cong}$, may arise from a better performance in congruent trials (i.e., greater facilitation as indexed by faster RTs), worse performance in incongruent trials (higher interference as indexed by slower RTs), or both. All these effects would support the hypotheses but it is behind the scope of the paper to investigate which specific mechanism may take place. Therefore, the dependent variable will be calculated as the difference in reaction times (RTs) between the two trial types, as our main research question concerns a modulation of performance. Clearly, this means that from the main analyses we will not be able to say whether the effects are explained by decreased RTs on congruent trials rather than by increased RTs in incongruent trials. Further analyses on the contribution of the different trial types may be left to explorations. We have clarified this point at p. 4, 8 and updated the Study Design Template at p. 12-15.

1C. Feasibility of methodology and analysis. The analyses are not sufficiently clear at present. Here are some required improvements:

- What is your approach to outlying data points (at the trial and at the participant level)?

R: Thank you for pointing this out, we have now specified more clearly the exclusion criteria both at the trial level and at the participant level. In the VTSC task, in each participant we will exclude trials exceeding +- 2 standard deviations within the same trial type; this will be done after transforming the data to obtain a normal distribution (p. 9). In the 2-PDT and cmPAS no trials will be excluded, according to the previous study (Zazio et al., 2019). At the participant level, as already stated, participants will be excluded in case they do not complete experimental sessions; the stimulation intensity in any session exceeds 90% of the maximal stimulator output (MSO) and/or

performance in catch trials during the cm-PAS in any session is below 50%. Moreover, also in response to your following comment, we add two other exclusion criteria for the VTSC: participants will be excluded in case their performance in catch trials and/or in tactile only stimuli is below 50%. Finally, at the group level, participants will be excluded in case in any block the dependent variables (see following paragraphs for further details) exceed ± 2 standard deviations of the group mean. The updated version of exclusion criteria can be found at p. 6; we have also added a schematic representation to make it clearer (Figure 1).

- Please specify which analyses will be performed on the tactile acuity 'global performance' measures (d' and criterion).

R: Having multiple outcome measures to test our main hypotheses may be counterproductive, because then it will not be clear in which cases the hypotheses will be confirmed or not, as correctly pointed out also by Reviewer 3. Therefore, in the case of the VTSC task, we decided to leave out the analyses on *d'* global performance and response criterion, which are secondary in relation to our research question.

- The dependent variable for the VTSC task is specified as the difference between incongruent and congruent trials. However, you are also measuring tactile-only trials. Please explain how the tactile-only trials will be incorporated in this analysis, as your analysis plan only specifies 'VTSC measures'. Please also specify follow-up analyses if an effect of cm-PAS is found on the difference between incongruent and congruent trials: as noted under 1B above, the difference could be generated by changes to processing for the incongruent trials, the congruent trials, or both; how will you investigate this?

R: Thank you for your comment. Tactile only trials will not be included in the main analysis plan, in which we consider as a dependent variable the difference between incongruent and congruent trials ($\Delta RT_{incong-cong}$; see our response to your previous comment). We proposed to collect also tactile only trials as an additional exclusion criterion, to control that participants are actually responding according to the real tactile stimulus. We have added this information in the exclusion criteria paragraph at p. 6 and in the description of the VTSC (p. 9).

- The main analysis comparing the effect of cm-PAS across control and clinical groups compares the two groups at the two timepoints (pre and post cm-PAS) for the various dependent variables, on the cm-PAS 20ms condition only. Given that the cm-PAS 100ms condition is a crucial control condition, please add another factor of cm-PAS condition (20ms, 100ms) to the analysis. Please also specify the dependent variables more clearly, as per the points above.

R: Thank you for this suggestion, we agree that the control condition (ISI) is crucial to test our hypothesis. As also suggested by Reviewer 2, we included the factor ISI in the main analysis testing Hypothesis IV, in a 2x2x2 Group x Time x ISI mixed-design ANOVA. Moreover, we have now more clearly specified the dependent variables included in the analyses (i.e., sensory threshold in the cm-PAS and $\Delta RT_{incong-cong}$ in the VTSC task), also in the Study Design Template (p. 12-15).

- Please consider including session order as a variable to account for learning / carry-over effects.

R: It is true that session order may explain a portion of the variability of our results. However, given that sessions are fully counterbalanced, possible significant effects cannot be explained by session order; thus, it is not essential to test our main hypotheses. To avoid the risk of reducing statistical power by adding another factor to our main analysis, we suggest leaving it for possible exploratory post-hoc analyses.

1D. Methodological detail.

- More detail required please of where patients will be recruited from, and where controls will be recruited from. Will both groups be community samples, for example? Authors mention matching on gender and age – will this be on a case-control basis and if not, how will this be done? What about ethnicity of participants – the hand stimuli display a White hand so will participant ethnicity be matched across groups?

R: Thanks for this comment, we have now integrated the manuscript with missing information (p. 4). Participants will be mostly locals: the patients will be recruited by the Psychiatry Unit, and HCs by word of mouth from the general local population. BPD patients will be matched one-to-one with HCs for gender and age; for age we will have a tolerability of +- 2 years. Based on the multi-decade experience of the Psychiatry Unit, the presence of patients from non-Caucasian is extremely low (about 3%). Therefore, we will not match for ethnicity.

- Which subscale(s) of the QCAE were used to measure cognitive empathy in the pilot data (page 4, point i) under sample size)? And which will be used in the main study (also which IRI subscales will be the focus of analysis)?

R: For the QCAE, both for the pilot data and for the main study, we consider the mean score given by the subscales 'perspective taking' and 'online simulation'. According to Reviewer's 3 first comment, we decided to leave out the administration of the IRI questionnaire. We have now specified this point at p. 7.

- Sample size estimation for the VTSC measure – I appreciate that the effect of cm-PAS on this measure has not previously been tested, making it difficult to estimate the likely effect size. The authors have therefore based sample size estimate on the visual acuity task. Please provide some indication of the relative variation in performance across the visual acuity and VTSC tasks or some other comparison between these tasks to convince readers that the effect of cm-PAS on the VTSC task is likely to be of the same order of magnitude as its effect on the visual acuity task.

R: Thank you for your comment. While the modulation of tactile acuity after the cm-PAS has been already tested in a previous study (Zazio et al., NeuroImage 2019), there are no studies which have investigated the modulation of performance in the VTSC after a PAS protocol. To the best of our knowledge, the closest study is the one by Bolognini et al. Curr Bio 2014, which employed a similar version of the VTSC task and an online TMS protocol over S1 to induce a modulation of performance. In the different experiments, the effect size observed for a significant interaction after a rm-ANOVAs is within a range of *ηp2* = 0.25 - 0.31. The effect size reported in Zazio et al., NeuroImage 2019 for the condition of interest on tactile acuity is ηp2 = 0.3. Therefore, it seems reasonable to consider a similar effect size for the power calculation in the two tasks. We have clarified this point at p. 5 and in the Study Design Template (p.12-15).

- Pharmacological treatments for BPD are likely to mean many potential participants in this group have contra-indications for TMS. Please comment on how representative your eventual sample is of patients with BPD in general.

R: Pharmacological treatments do not represent an exclusion criterion per se, according to the latest TMS guidelines (Rossi et al. 2021). For extreme caution, participants will not be included in the study only in case they take pharmacological treatments that decrease the epileptic threshold (e.g., clozapine, bupropion). Up to now, in the previous experience of the Research Unit of Psychiatry, only 3 out of 200 BPD (1.5%) patients were taking one of these two treatments. Based on these data, we believe that the sample we will test will be representative of BPD patients in general. We added this information at p. 4.

- Figure 1 suggests ISIs of 20 and 150ms, text suggests 20 and 100ms – which is correct?

R: Thank you for pointing out this typo; 100 ms is the correct one, we have now corrected it in Figure 2.

- With 11 participants per cm-PAS session order, task order cannot be fully counterbalanced – consider increasing sample size to 24 per group to allow full counterbalancing of session order and task order?

R: This is true. According to your suggestion, we have now increased the sample size to 24 participants per group to fully counterbalance also the task order.

- Will performance on catch trials on VTSC task be included as exclusion criterion? If so please specify performance cut-off for inclusion/exclusion, if not please indicate what the purpose of these catch trials is.

R: As outlined in our response to your previous comment we have included performance on catch trials as an exclusion criterion, as now specified at p. 6.

It would be desirable for the experimenter delivering the tactile stimulation in the tactile acuity task to be blinded as to participant group (control, BPD) and certainly as to cm-PAS condition (20ms, 100ms). Please confirm whether this will be the case.

R: Thank you for this suggestion. While it may be difficult that the experimenters are blind to the group, as they are both involved in the recruitment phase, having the experimenter that will deliver the 2-PDT blind to the cm-PAS condition is feasible and will increase the rigor of the study. We have specified this point at p. 9.

1E. Outcome-neutral conditions. The effect of cm-PAS in healthy controls is included as a positive control.

R: That's correct.

REVIEWER 2

In this project about tactile mirror system in BPD patients, Zazio and colleagues are going to collect data from 44 participants (22 BPD patients, 22 HC subjects). Participants will be asked to perform a tactile acuity task (2-PDT) and a visuo-tactile spatial congruity task (VTSC) twice,

before and after cmPAS protocol. In the cmPAS paradigm, participants will be provided with a single-pulse TMS over S1 while watching videos of virtual hand being touched. This procedure should lead to an increment of the tactile mirror capabilities in HC that would manifest with a possible better/worse performance in 2-PDT and VTSC tasks, respectively, in the second session than in the first session. Authors, however, do not expect the same plasticity in the tactile mirror system due to cmPAS in BPD patients. This expectation would be justified by the reduced empathic abilities of BPD patients, already proven in the literature. I have some minor questions about the paradigms (which could be useful for future research) and statistical analysis which I listed below. However, I think that authors provided a well-structured design both in terms of validity of research questions and hypothesis, experimental procedure and statistical planning. Therefore, my opinion is more than positive.

R: Thank you for your comments and your positive evaluation of our manuscript.

1. In the VTSC paradigm, the virtual hand is palm up and it receives the tactile stimulation on the palm. On the contrary, the real hand is palm down. Actually, I did not understand where the real hand receives the tactile input (palm or dorsum). You wrote palm in the text, but it seems dorsum in the figure. Please this issue should be clarify. Even if the real tactile stimulus is delivered on the palm (in line with the virtual hand), the real hand is palm down because of the manual response with the keyboard. Is this procedure the most ideal to obtain reliable results? Wouldn't be better to have the very same configuration of the virtual hand and real hand (e.g., both virtual hand and real hand palm up)?

R: We apologize if the location of tactile stimulus delivery during the VTSC was not clear; we have now modified Figure 3 (previously Figure 2) caption accordingly. The tactile stimulus will be on the palm, in the same location as the one depicted in the visual stimulus. However, the actual position of participants' hands and the one depicted on the screen will differ, as participants are asked to provide a response on a computer keyboard. While having the real and the virtual hands in the same position may boost the effects by inducing also an embodiment effect, we suggest that it may not be a *conditio sine qua non to* to observe mirror-like mechanisms. Indeed, in previous studies employing the VTSC, the real and virtual hands were not in the same position (Bolognini et al., Curr Bio 2013, 2014). We have now clarified this point at p. 9.

2. The response in the VTSC task is provided with the right hand; however, the response (right) hand receives the tactile stimulation in half of the trials. A pedal response (provided by foot) would have been better since no conflict arises between the stimulation and the response. Could you justify this choice? Moreover, it is not clear which are the fingers of the right hand used for the response.

R: Thank you for this comment. As regarding the hand position, we referred to previous studies using the VTSC task (Bolognini et al., Curr Bio 2013, 2014), in which the hand receiving the tactile stimulation is the same providing a response. Please note that participants are always asked to provide a response with the hand on which they received the tactile stimulus, irrespective of whether the visual touch stimulus is spatially congruent or incongruent trials. Therefore, the results cannot be explained by any influence of tactile stimulation on the response.

3. In the cmPAS paradigm, the fixation hand lasts more than 9 seconds. I see that this trick avoids a possible summation of excitatory inputs due to repeated pulses of TMS over time, but I think that participants could have a 'weird' visual experience due to the (very) long fixation and the (relatively) brief touch. Why did not you add an ITI between trials to make the fixation shorter?

R: A time interval of 10 seconds between a TMS pulse and the next one is indeed quite long. Classical PAS protocols typically use a fixed time interval between the stimuli pairs at a frequency which can be even lower (e.g., 0.05 Hz in Stefan et al., Brain 2000; Wolters et al., J Neurophys 2003) or equal (i.e., 0.1 Hz; Wolters et al., J Phys 2005) to the one proposed in the present study. Also 'modified' versions of the PAS targeting multisensory processing exploited a similar frequency of stimulation (see Guidali et al., Behav Brain Res 2021 for a review). Although a jittered version of the cm-PAS have been proposed (Maddaluno et al., Cortex 2020), it is important to note that the time interval could vary between 5, 10 or 15 seconds, so the inter-stimulus interval could be even longer between one TMS pulse and the next one. Here, we preferred to be consistent with most of the existing literature of PAS protocols, and specifically with our previous study on cm-PAS in which the effect was observed in three different experiments. Considering the well known inter- and intra-subjects variability of PAS effects (Guidali et al., Behav Brain Res 2021), we believe that keeping the same protocol's parameters from previous successful studies is a preferable approach for the first study applying the cm-PAS in a clinical population. Double-touch trials have been introduced especially to keep participants' attention on the presented stimuli during the cm-PAS, and to control for it (i.e., participants with a performance below 50% will be excluded from the sample).

4. In the cmPAS protocol, it is not clear how authors will find the targeted area S1 (if they have specific coordinates from the literature or if they follow other procedures). Maybe this information is not needed in this first version of the manuscript.

R: To identify S1 area, we will follow previously described procedures: starting from APB motor hotspot, defined as the highest and most reliable motor-evoked potentials with the same TMS intensity, we will move the coil 2 cm laterally and 0.5 cm posteriorly, according to Holmes & Tamè, J Neurophys 2019 and Holmes et al., J Neurophys 2019. This information can be found in the 'Transcranial magnetic stimulation (TMS)' paragraph at p. 8.

5. Why did not you opt for a 2x2x2 design in the final ANOVA with between-subj factor Group and within-subj factors Time and ISI? I do not understand your choice to split the final analysis in two parts, depending on the results of the first.

R: Thank you for this comment. As also outlined in response to Reviewer 1, we have now included the factor ISI in the main analysis testing Hypothesis IV, in a 2x2x2 Group x Time x ISI mixed-design ANOVA. We have updated the manuscript at p. 11 and in the Study Design Template (p. 12-15).

6. To make your final results more robust, you could support the frequentist statistics with bayesian statistics.

R: Thanks for this suggestion. As also suggested by Reviewer 3, we have included Bayesian statistics. Therefore interpretation of results will be based on frequentist statistics but we will now support the results with equivalent tests using Bayesian statistics. We updated the manuscript at p. 11 and the Research Design Table at p. 12-15.

REVIEWER 3 (Zoltan Dienes)

The authors have written clearly on the background to their study, its methods and how they will analyze the results. They have also considered the power for each effect separately. However, their power calculations are based on the mean estimates of previous work, which means they have not thereby controlled their error rates for missing effects somewhat smaller than this but still practically or theoretically interesting. That is my main point for the authors to address.

R: Thank you for your positive comments and suggestions.

Hypothesis 1:

Two tests will be conducted. Under what conditions will it be asserted the groups differ, and under what conditions that they are the same? If either f the two tests being significant results in the conclusion of a difference, then a correction of multiple testing is needed. How will sameness be inferred? Power was calculated for an effect of about d = 1. But presumably a known population difference of d = 0.5 would not count as the groups being sufficiently similar that they have practically the same empathy.

My personal approach would be to use a Bayes factor, with H1 modeled as half-normal with a SD (scale factor) of the raw difference in questionnaire units that has been found between the groups before, because this is the predicted effect size. However for power one needs something else; namely the effect size we just don't want to miss out on detecting. Otherwise we have no justification for asserting H0, should there be a non-significant result - because we have not controlled Type II error rate for effects of interest (see https://doi.org/10.1525/collabra.28202). One heuristic in that reference that may be useful, is using the lower end of a CI (maybe an 80% CI) on previous tests of BPD-HC differences as the roughly smallest plausible effect that may still be of interest.

R: Thank you for pointing out that we overlooked this issue for Hypothesis 1. We therefore suggest to consider the QCAE only to test Hypothesis 1, for two main reasons. First, the sample size was estimated based on preliminary data from this questionnaire; Second, the QCAE has been proposed to overcome some intrinsic limitations of the IRI, both from psychometric (Chrysikou & Thompson, Assessment 2016) and theoretical (Michaels et al., Psych Res 2014) points of view. Considering that we don't have a priori hypothesis on the IRI that cannot be addressed by the QCAE, we will not administer it. This has been updated in the manuscript at p. 7 and in the Research Design Table at p. 12-15. Please see our response to your following comment regarding the Bayesian approach.

Hypothesis 2:

The issue of having enough power to assert there is no interaction arises here as well. Admittedly the authors do not claim they will assert there is no interaction, only that a non-significant result will not confirm that one exists. But that is an inferentially weak position to be in: It would be good to have evidence whether the effect does or does not hold. Again one could either use a Bayes factor, or else use the lower end of a CI to inform the power calculation.

R: Thanks for pointing this out. As specified also in response to Reviewer 2, we will support the results obtained with the frequentist approach with Bayesian statistics, so that we can provide evidence for no effect. We updated the manuscript at p. 11 and the Research Design Table at p. 12-15.

Hypothesis 4:

Rather than predicting simply a medium size effect, which does not have a scientific justification as such, it may be best to think in terms of what difference in (e.g.) d' would be sufficiently meaningful.

R: We see your point. However, in the present work, tactile acuity is exploited as a proxy to test plastic mechanisms within S1, rather than representing a meaningful target of modulatory effects in BPD. Therefore, identifying a "meaningful difference" in tactile acuity between HC and BPD may not be straightforward. Since both approaches can be considered quite arbitrary, we preferred to predict a medium effect size, which is a well accepted approach in the literature to calculate the sample size in the absence of preliminary or previously published data.

Other points:

abstract:"Here, we take advantage of a cross-modal PAS (cm-PAS) protocol" define PAS.

R: Done, thanks.

p 7

" after transforming the data in case of non-normality of the distribution."

How exactly will non-normality be established? Some details are given later, but not a precise decision rule. What transformations will be used? Provide a set of if-then rules to tie down analytic flexibility.

R: Thank you for highlighting this lack of clarity in our statistical approach description. We preferred not to rely on the results of formal normality tests (e.g., Shapiro-Wilk test) because their statistical power is low for small sample size (so failure to reject cannot be used to conclude normality) and graphical/numerical methods are indeed preferable, although they do not provide a threshold for establishing whether a distribution is

normal or not (Mohd Razali & Bee Wah, J Stat Modeling and Analytics 2011). We have now better specified how normality will be evaluated at p. 11.

" The sensory threshold will be estimated by fitting a logistic function to d' values (transformed to fit in a range between 0 and 1;" State how the fitting is done.

R: Thanks for highlighting that this was unclear. The fitting will be done in R using the 'fitting generalized linear models" (*glm* function, binomial family), and then we will consider as sensory threshold the distance value (on the x axis) corresponding to 50% performance (on the y axis), as now reported at p. 10.