ROUND 2

Chris Chambers

The three reviewers who assessed your initial Stage 1 submission returned to evaluate the revised manuscript, and the good news is that the manuscript is getting closer to achieving Stage 1 in-principle acceptance (IPA). As you can see, however, there remain some significant issues to address in the review by Zoltan Dienes to ensure that the hypotheses, sampling plans, and a priori evidence thresholds are fully defined and justified. As Zoltan is a member of the PCI RR Managing Board, feel free to contact him directly if you need any additional assistance addressing these points.

I look forward to receiving your revised manuscript, which will be re-evaluated by Zoltan and the Managing Board before issuing a final Stage 1 decision.

R: Thanks for your encouraging feedback. Please find our response below together with the changes in the manuscript highlighted in bold.

Looking forward to hearing from you at your earliest convenience.

Best Regards,

Agnese Zazio (on behalf of all authors)

REVIEWER 1

The authors have addressed my previous points with the exception of the following: Exclusion criteria (page 6, figure 1): for group level exclusions, text says +/-2SD, figure 1 says 2.5 SD – please clarify which is correct.

R: Thanks for noticing this. We have now corrected the text with 2.5 SD.

REVIEWER 2

Thanks to the authors for the modifications done to the manuscript and taking into account my comments. I have no other comment to make.

R: Thank you.

REVIEWER 3 Zoltan Dienes

They key point of my review - making sure the size of effects in relevant predictions are scientifically justfied - has not been addressed. The authors have added Bayes factors - though not specified nor justified the model of H1 used in the Bayes factors - but then still used frequentist power analysis to justify N. Further, they have used a decision threshold of 1 for the Bayes factor, allowing decisions based on virtually no evidence. Finally, if decisions

are to be with respect to the Bayes factor (according to the Design Template), it is not clear what if anything will be done with the frequentist statistics (which are said to drive the interpretation, according to the text) - there is scope for plenty of inferential flexiibility here. One inferential system must be fully specified, including how conclusions will follow from that system.

R: We apologize if we unclearly described how we aim to use Bayes factors in our manuscript. We agree that adding the Bayes factors in the Study Design Template together with Frequentist statistics was misleading and prone to inferential flexibility, as it was unclear which of the two would drive the interpretation in case of inconsistencies; we have therefore removed the bayesian statistics from the Study Design Template. Nevertheless, we suggest incorporating Bayes factors in the report of the results, to provide a more complete picture of our findings, but to use the frequentist approach to guide the interpretation and, thus, the sample size calculation.

I am happy for the authors to go over to Bayes factors if they wish (in fact I think it will be easier, but I leave it to them to judge that). But they need

1) To estimate an expected sample size based on the properties of the Bayes factor they use, and not basedon frequentist power analysis, see e.g. for a very quick overview of how to do this see https://www.youtube.com/watch?v=10Lsm_o_GRg

For the above estimation of N, the BFs that will be used to test predictions must be used. And for each of those BFs, the predictions of the theory need to be modeled - the model of H1. That is, my point about power requiring justification of the effect size used is not avoided by using Bayes factors. Let me be clear about why:

To test a theory via one of its predictions, the prediction must be justified as relevant to the theory. Obviously, if it is not relevant, falsifying the prediction does not count against the theory. Example i) if the prediction is for an effect, and the exact size of the effect found in a previous study is used for power, the prediction is modeled effectively as: The minimally interesting effect is the effect in the previous study; that is, missing any effet smaller than this is OK, as such effects are too small to be interesting. But this final conclusion is typcailly false. Typically effects smaller than that found in a preivous study are interesting. That is why PCI RR says in the Guide for Authors: "Power analysis should be based on the lowest available or meaningful estimate of the effect size." ii) If an arbitrary number like d = 0.5 is used for the prediction (arbitrary in many ways, including the dependence of the population d on the number of trials in the study), then the prediction is arbitrarily related to the theory (just as the number of trials used is arbitrarily related to truth of a theory), and falsifying the prediction does not count against the theory. (To address this argument, the argument in itself must be addressed: Claims about what standard practice is do not address the argument.)

2) To specify the models of H1. A Bayes factors pits a model of H1 against a model of H0. For the BF to be relevant tothe theory, the model of H1 should eb relevant to modeling the predictions of the theory. Modeling H1 can be thought of as boiling down to estimating a roughly predicted size of effect (contrast power).

So how to model each prediction? Some suggestions to consider

Hypothesis I: This is straightforward as relevant previous study has been identified using the same proposed questionnaire between groups very similar as will be sued for the current study. As a roughly expected effect size is what is relevant, the previous different in scale Likert units could be used as the SD (scale factor) of a half-normal. See replication heuristic of https://doi.org/10.1525/collabra.28202

Hypothesis II Similarly there is a previous study using the same apradign; use the raw effect size as the SD of a half-normal.

Hypothesis III: Use the room to move heuristic of https://doi.org/10.1177/2515245919876960

Hypothesis IV: Use the room to mvoe heuristic of https://doi.org/10.1177/2515245919876960

R: We appreciate your comments and the approach you suggested, which we thoroughly took into consideration. Related to sample size calculation, we acknowledge that different approaches exist and the debate on the best practices is still open. We see your point when you say that when the exact effect size of a previous study is used for power, "missing any effet smaller than this is OK [...], but typically effects smaller than that found in a preivous study are interesting [...]". On the other hand, as you reported, PCI guidelines say "Power analysis should be based on the lowest available or meaningful estimate of the effect size". The approach we suggested based on frequency statistics falls in the first case, i.e., we based power analysis on the lowest available effect size, whenever possible. In the only case we could not base our sample size estimation on previous data (i.e., HP IV), we used the "medium" effect size, which is much smaller than effect sizes observed in the field, in particular for the other hypotheses (e.g., please see the "Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis" for HP III). Moreover, HP IV is about the difference between BPD patients and HC in the effects of the cm-PAS in the plasticity mechanism of the tactile mirror system: while it is true that small effects can be interesting, in this case a small effect would suggest that alterations of the tactile mirror system are not that crucial for borderline personality disorder.

In brief, although we cannot fully rule out the possibility that we may miss smaller but potentially interesting effects, we believe that our approach based on frequentist statistics should not be considered as flawed. We have now added a justification of the approach we used in the paragraph on sample size estimation, mentioning its potential limitations, at p. 5.

More analytic flexibility is introduced by not having scripted normality checks. I actually agree that the standard significance tests of normality are pretty much useless and checking by eye can be better. But that leaves analytic flexibility. In fact, the authors propose to use robust t-tests (trimmed, yuen t tests). In general there is no reason to think non-parametric tests produce more relevant results than robust t-tests, so I wold suggest simply using the robust t-tests (performing the equivalent Bayes factors by using the robust means and SEs that go into the robust t-test, then using a BF calculator such as https://harry-tattan-

birch.shinyapps.io/bayes-factor-calculator/). The authors worry about robust ANOVAs. But in fact the crucial test in every case is one degree of freedom and thus the relevant contrast reduces to a t-test. That is practically, find relevant differences or differences of differences for repeated measures variables; then perform a robust two group t-test on the difference of diferences (etc); from the robust t-test extract the robust SE of the effect and the robust estimate of the effect and enter these into the BF calculator.

If the authors have questions on the above they are welcome to contact me, if they wish.

R: Thanks for pointing this out, we agree that a visual evaluation to establish whether a distribution is normal or not leaves analytic flexibility. On the other hand, we would avoid reducing all the statistics to robust t-tests, as in the previous round of revisions we included a 2x2x2 ANOVA (instead of two 2x2 ANOVAs) in response to both the other Reviewers' suggestions; moreover, the ANOVA would allow us to test for main effects and interactions in a single model. All considered, although we agree that normality tests present some limitations, they provide an objective way to make decisions on statistical tests. Therefore, we suggest to perform Shapiro-Wilk tests and proceed with the data transformations whenever results significantly deviate from normality. Then, we will use the Cullen-Frey graphs to choose the transformation that brings the data closest to normality. This approach will avoid analytic flexibility. We have now updated the manuscript at p. 11.