ROUND 3

Chris Chambers

I consulted with Zoltan Dienes again and most issues are now settled. You will see that there are two remaining points to address concerning the selection of the smallest effect size of interest and the statistical testing procedure. Please consider these points carefully. Provided you are able to address these issues in a final revision, we should be able to proceed with Stage 1 in-principle acceptance without requiring further in-depth review.

R: Thank you for your feedback; we have carefully addressed the remaining issues raised by Zoltan Dienes, as outlined in detail below. Please find changes in the manuscript highlighted in bold. We hope that you will find the manuscript suitable for Stage 1-in principle acceptance.

Looking forward to hearing from you at your earliest convenience.

Best Regards,

Agnese Zazio (on behalf of all authors)

REVIEWER 3 Zoltan Dienes

The authors have clarified they will base their decisions on one system (frequentist hypothesis testing) to be clear how they have tied their inferential hands; which hand-tying they likewise do for assumption testing by being clear about how they will proceed with testing normality. (Using power in this way does not make best use of all data once it is in; but that is the authors' choice.) BFs will be reported for information only. So this deals with a key issue. But there remain a couple of points:

1) The use of power to determine N means they need to justify a roughly smallest effect of interest. The authors say "we based power analysis on the lowest available effect size, whenever possible." But the paper itself lists a single past study for each test and uses the effect size from that past study. Technically, this may then be the smallest available effect because there is just one. But then one has not plausibly controlled Type II errors so as not to miss very interesting effects. My main concern is respecting the spirit of the point; but here is a particular suggestion. Following ideas in the paper I previously referred to, https://doi.org/10.1525/collabra.28202 they could put a 80% CI on the one previous most relevant study, and use the bottom limit of the CI as an estimate of a smallish effect that is just plausible, and so long as it is interesting, that could form the basis of the power analyses.
R: We have carefully considered the method you suggested and applied it to our calculations for the sample size. For example, in Hypothesis I, considering the lower bound of 80% confidence interval of the effect size reported with the same levels of power (0.9) and alpha (0.02), results in 70 participants per group, more than three times the one obtained with the original effect size (i.e., 21). We have discussed running a

study with at least this sample with the Unit of Psychiatry, and unfortunately it is not feasible within our institution. In general, we would like to point out that we respected the PCI requirements about having a dedicated sample size calculation for each of the hypotheses in the study, and also about the levels of power and alpha required by a few PCI-friendly journals, while there's no specific requirement for the procedure of sample size calculation. On the other hand, we are aware of the risks of running underpowered studies. For this reason we suggest to make it explicit in the introduction that in the present study we are looking for strong effects, and potentially smaller but still interesting effects may not be detected (p. 4 and 5). Moreover, our interpretation of non-significant results will be tempered.

2) There are arguments for why it is better to use robust tests from the beginning rather than doing the two-step "significance test of assumptions -> choose test" procedure (e.g. Field & Wilcox, 2017 https://doi.org/10.1016/j.brat.2017.05.013). The authors are aware of these issues, but suggest because they are dealing with 2X2X2 effects, Yuen robust t-tests are ruled out. In fact, as far as I could tell, all crucial tests in the Design table involve either a repeated measures main effect or interaction on HC; or the difference in such an interaction between the two groups. While not all terms of the 2X2X2 ANOVA can be easily run as a Yuen t-test, all terms that involve an interaction with group can be; and all purely repeated measure interactions involving one group can be tested with a robust one-sample t-test. As I say, as far as I can make out, that applies to all crucial tests. Since all other tests are exploratory (pre-registered conclusions from them will not be drawn), they should in any case be reported in a separate section. This opens back up the option of being robust from the start; and this consierably simplifies the pre-registration. I leave this to the authors' judgment.
R: We see your point, and we agree that in some cases the approach you suggested is useful to avoid the limitations of normality testing required for the analysis of variance. However, we are also aware that every approach has its own limitations. In the case of the present study, we disagree that replacing the ANOVAs with robust t-tests represents the best approach: As typically happens when considering difference values some important information is lost and the results are more difficult to interpret. Especially in the 2x2x2 ANOVA of Hypothesis IV, for the triple interaction (Time X ISI X Group) we would need to calculate the difference between post-pre, then the difference between ISI-20 and ISI-100, and then use the obtained values to compare the two groups. In the case of a significant difference between groups, the interpretation is not straightforward, as we would not know which factor drives the difference. While it is true that we may add the ANOVAs as exploratory analyses, we believe that this approach would negatively affect the readability of the manuscript without providing a significant advantage.