Response to Reviewers

Recommender

1) I have received reviews from all three experts from the first round of reviews. All three are positive and once again, I completely share their overall positive evaluation, both regarding the proposed study and the revisions. Only a small number of remaining/additional questions have been raised, and I'm looking forward to your point-by-point response. I would only add the small observation that the formulation of hypothesis 3 under "final hypotheses" could be changed to specify the direction of the effect, and I wonder whether the word "significant" can/should be removed from the statistical hypotheses.

   Author Response: Thank you for the positive evaluation of our manuscript. In response to your points, the hypotheses have been altered to make directionality clear, and all references to "significant" findings removed.

Reviewer 1

2) "While it is certainly true that to go through on an individual trial basis would be an inappropriate way to analyse such data and would certainly increase the risk of bias, Clarke et al. (2014) observed this effect at the group level, suggesting there is sound evidence for this argument." I didn't understand this sentence, in relation to the counterarguments involving effects of ABM (which otherwise now seem very well described!). As I understand it, the criticism of Kruijt etc *is* about the interpretation of the pattern of results over studies (which is "at the group level" if I understand the phrase here correctly). The issue is that this pattern involves a kind of indirect cherry-picking - i.e., to select studies for an effect on bias at post-test could be to select studies for an effect on (e.g., a clinical) outcome with *any* association with the bias, without this implying specifically that causal relationship that runs from bias to outcome. That's merely one possible interpretation - but, e.g., a sceptical observer could equally posit the possibility that p-hacking will tend to generate pairs of false positives for both bias and outcome that tend to occur together in particular sets of studies; or, perhaps improvements in outcome over time tends to cause changes in bias over time, even if the effect of ABM on outcome was a false positive, so selecting studies on change in bias means implicitly picking out false positives on outcome.

   However, I feel like the literature has been presented clearly and sufficiently, so making this argument is up to the authors - whether or not it's a good or bad argument can be judged by readers. I'd just suggest that perhaps the issue is best explicitly described in terms of a high degree of uncertainty and speculation given the available (lack of) evidence - it could well be possible that the pattern of results indeed reflects only some ABM experiments causing a change in bias, and this factor causing a change in outcome; but the pattern of results doesn't provide evidence for that particular interpretation of it over other possibilities.

   Author Response: Thank you for your thoughts on this issue. You are quite right, and we have altered our discussion of this on P.8-9 to reflect the uncertainty in this debate.

3) "We agree with the response raised by Parsons (2018), whom argues that" - "whom" should be "who".

Author Response: Thank you for pointing this out, it has been corrected.

4) "However, in line with advice from a discussion with Professor Zoltan Dienes, we will retain our final analytical decision threshold at BF >= 3 as evidence for H1, and BF <= 1/3 as evidence for H0. This is because if you have the same threshold on your stopping rule as you have on the analytical decision threshold, then the Robustness Regions reported will show no robustness (essentially by design) as you stopped data collection the moment it reached that point." I wasn't sure I understood the argument here. Does "final" in "final analytical decision threshold" mean the threshold used is the maximum sample size is reached? If the stopping criteria are 30 and 1/6, then the thresholds of 3 and 1/3 will be irrelevant except in the case the maximum sample size is reached, but I'm not sure what the problem with "robustness" mentioned in the response would be to maintain the 30 and 1/6. However, as above, if the authors are comfortable this is correct and will be clear enough in the text to readers, as mentioned I'm not an expert; otherwise it might be helpful to try to clarify the rationale.

Author Response: We are striving for a high level of evidence and hence set the stopping rule to be 1/6th and 30 on the key outcome variable. So, for our central analysis it is correct that, provided we don't stop due to reaching our maximum participant number, the evidence would support a decision threshold at those levels. However, our stated decision threshold of 3 and 1/3 will be applied to all analyses, including any exploratory analyses etc. While we agree that setting the stopping rule for H1 to be 30 makes sense in the current context, many readers would find it odd if we dismissed as insensitive results reaching say 25. We therefore apply a decision threshold (the Bayes factor value at which we will report there being evidence) at the standard level for moderate evidence (3 and 1/3) while reporting both the Bayes factor (telling the reader the actual strength of evidence achieved in each instance) and the robustness region (telling the reader how much variation in the scale factor we selected - which we always seek to justify as fully as possible - would result in the same conclusion). Reporting things in this way enables the reader to check at a glance both whether the level of evidence meets their unique preference, and examine if their preferred scale factor (perhaps derived from a different study they are familiar with) would result in the same outcome.

5) "For all Bayes Factors we will adopt the conventional thresholds of values greater than 3 indicating evidence for the alternate hypothesis and values less than 1/3rd indicating evidence for the null." and "Robustness regions will be reported as: RRconclusion [x1, x2], where x1 is the smallest and x2 is the largest SD that gives the same conclusion: B < 1/3; 1/3 < B < 3; B > 3." Possibly related to the above, is this still correct / will this be clear given the proposed changes to 30 and /6?

Author Response: Yes, as outlined in our response to Point 4, this remains correct.

6) "In using the procedure detailed by Palfi & Dienes (2019, Version 3, p. 15), it was determined that given a long-term relative frequency of good enough evidence of 50%, the proposed sample size allows for a discriminating Bayes factor (B > 30 if H1 is true, and a B < 1/6 if H0 is true)." Is this still correct, since the numbers in brackets changed while the rest of the sentence didn't?

   Author Response: Yes, this remains correct, our maximum participant number of 200 meets the requirement for the increased threshold (as outlined in Footnote 4, P. 16).

Reviewer 2

7) The authors have answered my comments in detail and satisfactorily - thank you very much. In my view, this is exciting and methodological sound study. I am very much looking forward to the results!

   Author Response: Thank you for the positive appraisal – we look forward to sharing the results!

Reviewer 3

8) I have reviewed the responses made by the authors with regard to my comments. I am overall happy with the responses they gave. One concern remains with regards to the data analysis plan. I understand the considerations of the authors, however, the ABM field would benefit greatly from taking into account the many random factors that come into play and that can have quite a substantial effect on the outcomes. I do however agree with the added value of the Bayesian approach and can see that not all limitations in a field can be addressed in one study. I would recommend acceptance of the stage 1 report at this point.

   Author Response: Thank you, we are glad you find our report to be of IPA quality.