Response to Reviewers

Recommender

1) I have now received the detailed and helpful evaluations of three experts. They all welcome the proposed replication study as a relevant contribution to the field of ABM research. I share their overall positive evaluation and believe that this submission is a promising candidate for eventual Stage 1 in-principle acceptance. I will not attempt to reiterate all of the detailed and constructive points that have been raised, especially as the reviewers point out specific ways in which these concerns can be addressed. I would only like to highlight a few issues that appear particularly important.

   Author Response: Thank you, we are glad you find our work to be a strong candidate for Stage 1 IPA.

2) First, with respect to the adequacy of the sampling plan, I agree with the observation by Dr. Gladwin that the combination of low minimum N (n_min=11 per condition) and a lenient stopping rule (BF>=3) may be perceived as concerning. With these parameters, the risk of false positive evidence appears to be avoidably high, while the achieved evidential standard is only weak to moderate. Regarding this issue, Schönbrodt and Wagenmakers (2018) write: "False positive evidence happens when the H1 boundary is hit prematurely although H0 is true. As most misleading evidence happens at early terminations of a sequential design, the FPE rate can be reduced by increasing n_min (say, n_min = 40). Furthermore, the FPE rate can be reduced by a high H1 threshold (say, BF10>=30). With an equally strong threshold for H0 (1/30), however, the expected sample size can easily go into thousands under H0 (Schönbrodt et al. 2015). To avoid such a protraction, the researcher may set a lenient H0 threshold of BF10<1/6". Thus, I encourage you to carefully revisit their sampling plan according to these considerations.

   Author Response: Thank you for your helpful advice regarding the sampling plan. In line with your advice, we will now collect n_min of 40 participants and will set our stopping rule at BF >= 30, or <= 1/6 (as reflected in the 'Participants' section of the manuscript, see P. 15-16). 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.

3) Second, regarding the analysis plan, the reviewers also noted that some clarification is needed regarding the precise statistical methods and the mapping between hypothesis and statistical tests. Other points of note include potential limitations of the operationalization of demand characteristics, and that the presentation of the literature underpinning the research question can be strengthened further. You may also find the suggestion helpful to complement the sampling and analytical approach with the frequentist analyses used by Hazen et al. (2009) and/or power analysis for smallest effect size of interest (e.g. to determine n_min).

Author Response: Thank you for summarising these key points – please see our responses to Points 4, 9, 10, and 15, which seem most relevant in addressing these points. However, in terms of the suggestion to include the original Frequentist analyses conducted by Hazen et al. (2009), we wanted to highlight our response. All of our inferences will be based on rigorously calculated Bayes Factors following the guidance provided by Dienes (2021), but for each analysis where we report a Bayes Factor, we will also report the corresponding t and p-values. However, after careful consideration, we propose not to report additional tests (such as the original frequentist analyses) from which no inferences will follow, as to do so would be to invite non-pre-registered inferences that would be inappropriate in a registered report.

Reviewer 1

4) Under this heading, I primarily have some concerns about the presentation of the literature underpinning the research question. In terms of the literature, the debate around ABM seems to deemphasize arguments from one side, expressed in particular in: Kruijt & Carlbring (2018), "Processing confusing procedures in the recent re-analysis of a cognitive bias modification meta-analysis", https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/processing-confusing-procedures-in-the-recent-reanalysis-of-a-cognitive-bias-modification-metaanalysis/43E057467A6217353E3297B31B18A1E2, and Cristea (2018), "Author's reply", https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/authors-reply/6BEE25F8DBF57BC6DBD0A026A16E5762.

   E.g., from Cristea's reply: "Yet a larger and more crucial problem relies in the central claim of Grafton et al, echoed by many leading CBM advocates: the effectiveness of these interventions should only be weighed if they successfully modified bias. Kruijt & Carlbring adeptly liken this to familiar arguments for homeopathy. However, it also reflects a fundamental misunderstanding of how causal inferences and confounding function in a randomised design. Identifying the trials in which both bias and outcomes were successfully changed is only possible post hoc, as these are both outcomes measured after randomisation; reverse engineering the connection between the two is subject to confounding. Bias and symptom outcomes are usually measured at the same time points in the trial, thus making it impossible to establish temporal precedence.Reference Kazdin4 Circularity of effects, reverse causality (i.e. bias change causes symptom change or vice versa) and the distinct possibility of third variable effects (i.e. another variable causing both symptom and bias changes) further confound this relationship.Reference Kazdin4 For instance, trials where both bias and symptom outcomes were successfully modified could also be the ones with higher risk of bias, conducted by allegiant investigators, maximising demand characteristics or different in other, not immediately obvious, ways from trials where neither bias nor symptoms changed. Randomised controlled studies can only show whether an intervention to which participants were randomised has any effects on outcomes measured post-randomisation.Reference Kaptchuk5 Disentangling the precise components causally responsible for such effects is speculative and subject to confounding. To this point, randomised studies show CBM has a minute, unstable and mostly inexistent impact of any clinically relevant outcomes." While this is all in the

context of a debate with clearly varying opinions on the merits of different positions and analyses, it does seem to me important to accurately represent all sides and present any strengths of their arguments as well as possible.

Author Response: Thank you for highlighting these very important papers. We have added a paragraph (see P. 8) discussing the alternative perspective raised by Kruijt & Carlbring (2018) and Cristea (2018).

5) I'd additionally suggest that another elephant on the room that would be worth mentioning, especially given the advantages of the current approach of writing a registered report, is the replication crisis and the potential role of questionable research practices in general, to which ABM/CBM research hasn't necessarily been immune.

Author Response: Thank you for raising the need to call attention to the importance of registered reports and their role in the replication crisis. We have added a short paragraph discussing this in the introduction (P. 10, second paragraph).

6) However, also with an arguably fuller representation of the debates, I still think the research questions of the registered report remain scientifically valid.

Author Response: Thank you, we are glad you find our research questions to be overall valid.

7) I have no concerns with the hypothesis of an effect of the ABM training.

Author Response: Thank you, we are glad you find our main hypothesis to be sound.

8) The secondary hypothesis, on demand characteristics, seems only partly sound. The issue is how strong and one-to-one the auxiliary assumptions would have to be to work back from a possible null effect on the current measure of Phenomenological Control back to a conclusion on demand characteristics as envisioned, in particular, by Cristea et al. (2015).

Author Response: We think that we have addressed this concern in our response to point 10.

9) I am not an expert in Bayesian methods, so these comments are only intended as observations for consideration in case they're helpful.

First, I think a replication of Hazen et al. (2009)'s statistical approach would be very helpful to include, even if the author's specify their Bayesian approach takes precedence for their conclusions. If there's a discrepancy, say, a non-sigificant effect using significance testing but evidence considered supportive with the Bayesian

analysis, then I think readers might want to know and evaluate what could explain that.

Second, relatedly, I'd be concerned if the current method produced a sample that would be considered underpowered from other perspectives; in principle, as per the current method, this could potentially end up being N=23. This also perhaps relates to the Bayes Factor cut-offs proposed here (i.e., the analogue to the .05 p-value) of 3 and 1/3, which are only just past what would be considered "weak" and into a "moderate" range (see, e.g., van Doorn et al., 2021, The JASP guidelines for conducting and reporting a Bayesian analysis). It seems that the approach, dependning on how the first few dozen observations work out, might allow a "support-refute" decision that would easily be overstated given the evidence. E.g., from van Doorn et al. (2021), "The strength of evidence in the data is easy to overstate: a Bayes factor of 3 provides some support for one hypothesis over another, but should not warrant the confident all-or-none acceptance of that hypothesis."

Author Response: Thank you for your helpful suggestions regarding the planned analyses. With regards to your first point, all of our inferences will be based on rigorously calculated Bayes Factors following the guidance provided by Dienes (2021), but for each analysis where we report a Bayes Factor, we will also report the corresponding t and p-values. However, we propose not to report additional tests (such as the original frequentist analyses) from which no inferences will follow, as to do so would be to invite non-pre-registered inferences that would be inappropriate in a registered report. With regards to your second point, we agree that the originally proposed sampling plan required improvement. Therefore, in line with the suggestions made by yourself and our PCI-RR recommender, we will now collect n_min of 40 participants and will set our stopping rule at BF >= 30, or <= 1/6 (as now reflected in the 'Participants' section of the manuscript, see P. 15-16). 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.

10) As noted above, it doesn't seem like a null effect for the secondary hypothesis would be very meaningful, at the design/measures level; i.e., even if very strong Bayesian evidence for the null were found, this wouldn't address whether the one particular operationalization adequaltely represents the effect of demand characteristics. This potentially could be mitigated by creating a more meaningful test of demand characteristics, e.g., by including additional measures and concepts. Or, this test could be acknowledged to be quite weak and not to be overinterpreted. Maybe it would even be useful to take a more exploratory, qualitative view and use interviews asking participants about experiences related to demand characteristics.

Author Response: Thank you for raising this issue – we have added discussion of the limitations of the PC scale as a measure of demand characteristics into the

manuscript (see the added detail to the demand effects paragraph on P. 12-13). We agree that a null finding for a relation with PC should not be taken as strong evidence for the absence of demand characteristics, and have sought to address this by acknowledging the limitation of the PC analysis in the case of a null result (see the aforementioned sections of the revised manuscript). We have a PC database at the University for current students, meaning that when people take part in this research study, we already have participants' PC scores securely on file (with express permission!) and are therefore able to link PC scores to our final dataset without changing the original methodology of Hazen et al. (2009). The inclusion of further measures of demand characteristics, as well as the inclusion of an interview, would change the methodology of the original study and thus interfere with it being a replication. Furthermore, to interview participants about their experience of demand effects is often of limited use, as Orne (1962, p. 780) points out: "even if the experimenter makes an effort to elicit the subject's perception of the hypothesis of the experiment, he may have difficulty in obtaining a valid report because the subject as well as he himself has considerable interest in appearing naive."

Reviewer 2

11) In the proposed stage 1 replication study, the work of Hazen et al, 2009 will be directly replicated.

    In my view, it makes absolute sense to replicate ABM studies. In particular, I think that the variance in the findings is due to the fact that many researchers repeatedly change the experimental paradigms in such a way that comparability is more difficult (e.g., different presentation times, image sizes, designs, etc.). This is a point that the authors can also gladly include in their argumentation.

    Author Response: Thank you, we are glad you find our study to be of value. We have added some brief discussion of the variability in experimental paradigms on P. 5 and 8.

12) Delimitation and connection between attention and interpretation bias should be described in more detail in the theory. Concerning this aspect, it would be great if the authors would argumentatively present why it is worth to look more closely at only one bias and not at the connection, e.g. in the context of a combined cognitive bias hypothesis (Everaert et al., 2012)? (Everaert, J., Koster, E. H., & Derakshan, N. (2012). The combined cognitive bias hypothesis in depression. Clinical psychology review, 32(5), 413-424.)

    Author Response: Thank you for highlighting the need to discuss combined cognitive bias hypotheses. We have added a footnote on P. 4 discussing these in the context of the Hirsch & Mathews (2012) model that underpins the present research. Furthermore, we have provided additional justification here as to why the focus of the research will remain on attention bias.

13) It is somewhat irritating that the authors highlight in great detail and appropriately the previous meta-analytic effects on ABM in different disorders, but then propose a replication of a study in which precisely this distinction does not matter. Perhaps it would be sufficient for the authors to focus more on the results of GAS on pages 6 to 8.

Author Response: Thank you for raising this concern. While we do appreciate this concern, we feel it is important to highlight that the potentially reduced effects in samples relating to disorders such as SAD may result in an underestimate of the value of this intervention in GAD. It is clear from the literature review that the large number of previously reported null findings in ABM research seem to be driven by an oversaturation of studies performed on SAD samples, a community in which effect sizes generally seem to be smaller and often non-significant. However, the meta-analyses suggest that more generalised anxiety/worry samples seem to show more malleability of attention bias, and in turn symptomology. We feel this helps justify why a replication specifically looking at the effect on GAD is worthwhile (given that high worry is the central cognitive component of GAD, our intended sample represents a sub-clinical GAD sample). Furthermore, GAD is highly co-morbid with other anxiety diagnoses and, if found to replicate, using ABM to treat transdiagnostic worry across the anxiety disorders may have clinical value.

14) Could the authors give a direct example of pheomenological experiences? (S. 10)

Author Response: Thank you for the suggestion, we have added an example of a phenomenological experience on P. 12.

15) The design is very detailed and accurately presented. It would be helpful if the authors could present the central hypothesis again more clearly on p. 9, describe here which outcomes exactly confirm the hypotheses and also take up the hypothesis again explicitly in the context of the presentation of the Baysian stopping rule (p.10). It would also be helpful for the "mapping between hypothesis and statistical tests" if the authors would present the hypothetical predictions again in a formal-statistical way.

Author Response: Thank you for your recommendations. The final hypotheses have been added on P. 13, statistical hypotheses have been added on P. 14, and the hypotheses have also been presented in the context of the Bayesian stopping rule (P. 16).

16) For a better understanding, a figure presenting the trial procedure would be helpful.

Author Response: Thank you for your suggestion – we have added a figure presenting a typical trial procedure (see Figure 1).

17) Were the word pairs validated with respect to valence and arousal?

Author Response: Hazen et al. (2009) report that they used a subset of the word pairs used by MacLeod et al. (2002), and that these word pairs had been matched for length and familiarity. Upon closer inspection of the MacLeod et al. (2002, p. 109) paper, it emerges that the word pairs were also validated with respect to emotional valence – this is now reflected on P. 18. Given we take the word pairs from the same source, ours are similarly validated with respect to valence. Note that although we had to create the neutral-neutral word pairs ourselves from a subset of the neutral words listed by MacLeod et al. (2002), the authors report that all of these neutral words were close in emotional valence (with all neutral words having an average rating of between 4.6 and 5.9 on a 9-point emotional valence scale), meaning that any combination of the neutral words presented should be close in emotional valence. Unfortunately, we have no information regarding the emotional arousal of the words within each pair. However, in the spirit of a registered replication, we feel it is most important to seek to use the same materials as far as possible rather than generating new materials which may cause a disparity.

18) The planned analyses seem very appropriate and adequate to answer the research question.
Author Response: Thank you, we are glad you find our planned analyses to be appropriate.

19) The sample size is sufficiently planned - especially with regard to the hypothesis.

Author Response: Thank you, we are glad you find our sample plan to be sufficient.

20) By using Bayesian hypothesis testing, the authors will not infer evidence of absence from null results.

Author Response: Thank you, we agree that the use of Bayesian hypothesis testing overcomes this common analytical mishap encountered in Frequentist approaches.

21) There are already positive ethical votes for the study.

Author Response: Yes the study has full ethical approval and is ready to go!

Reviewer 3

22) The authors' pre-registered report describes a relevant replication in the ABM field with a valuable addition; namely addressing the demand effects within the laboratory setting. The report is well-written and incorporates a clear theoretical overview of the ABM literature. I have some concerns specifically pertaining the data analysis strategy that was chosen that should be addressed before the manuscript can be resubmitted. If this is addressed, the study will make a valuable addition to the literature.

Author Response: Thank you, we are glad you find our study to be valuable addition to the field, following small amendments of course.

23) The authors address attention bias for GAD. It would be helpful, especially for generalized anxiety, to give an example of attention bias for GAD.

Author Response: Thank you for your suggestion – we have added an example of attention bias in GAD (P. 3).

24) In the introduction the procedure of ABM is introduced, make clear that the target probe replacement is manipulated so that it more often replaces the neutral stimulus.

Author Response: This is now stated on P. 5, on the final line of the first full paragraph on this page.

25) Can the authors still add a reference for the study on demand effects in the lab on p. 5?

Author Response: Thank you for pointing out the need for clarity, the section in question is still referring to Cristea et al. (2015). This has been made clearer now (see P. 6).

26) In the replication of the study by Haazen et al., the authors choose to follow the choice for a composite of anxiety and depression as primary dependent variable. Even though comorbidity with depression is high for individuals with GAD, in a high worry, subclinical sample, this won't be relevant for all individuals and may obscure results. I wonder whether this composite outcome variable is also the standard in other ABM trials for general anxiety. I would like to see a short discussion on this in the introduction to help the reader place this choice adequately in the literature. I would suggest to at least also analyze these two constructs (anxiety and depression) separately (if necessary, in a supplementary file).

Author Response: Thank you for raising this issue. While a composite measure isn't necessarily a 'standard' choice in the literature, Hazen et al. (2009) used one in the interests of reducing the familywise error rate by reducing the number of individual statistical tests that would need to be performed (seeing as they'd only need to analyse one composite measure as opposed to three individual measures). Additionally, they hoped that the increased reliability and validity of findings afforded by doing this would increase their statistical power. Seeing as this is a replication of their method, we aim to replicate their analysis as closely as possible, which includes calculating and reporting their composite measure. However, as we are using a Bayesian approach we do not have the same concerns of the familywise error rate, and therefore we will also be running and reporting Bayesian analyses on each of the individual mental health measures (PSWQ, STAI-T, BDI; see final paragraph on P. 24, rolling into P. 25). Therefore, anxiety and depression will also be analysed

separately as you suggest. A brief footnote has been added in the introduction (see P. 11) briefly summarising the reasoning for including the composite measure.

27) I wonder about the role of baseline attention bias levels. This varies considerably in the literature (and probably specifically in a subclinical sample) and has also led to mixed results in the CBM field. It would thus make sense to at least control for this in the analyses.

Author Response: Given the random assignment to conditions, the likelihood of a substantial baseline difference in attention bias levels between the experimental and control group should be small. However, we do plan to report the Bayesian Credibility Interval on the difference between groups in attention bias (PDT) scores at pre-training to provide a clear estimate of the difference (see 'Preliminary Analyses' section of manuscript, P. 23).

28) My main concern is with the data-analysis part. It does not become entirely clear to me what specific analyses are being conducted. The authors describe their reasons for conducting Bayesian analyses instead of the original analyses, which are sound. However, which specific type of Bayesian analyses (e.g., based on ANOVA's, mixed effects models?) will be conducted and, with which type of program (e.g., how will the bayes factor be computed, with which program) – please clarify this.
I would suggest (if not already implied in the data-analysis section) to conduct mixed effects models considering the nestedness and random factors inherently present in dot-probe/ABM designs (e.g., trials nested within persons, training sessions nested within persons, random slope for stimuli etc.). Mixed effects models can also be conducted 'Bayesian style' (see the brms package by https://paul-buerkner.github.io/brms/, which is very user-friendly). Further, I would suggest, in the interest of replication, to conduct the original analysis of the Haazen et al. study as well to be able to make fair comparisons.

Author Response: Thank you for highlighting the need for clarity in the Planned Analysis section. To clarify, we are performing a series of Bayesian t-tests, which are being used as a direct substitute for the Frequentist t-tests performed by Hazen et al. (2009) – this has also been clarified in the manuscript (see start of 'Main Analyses' section, P. 23). In doing so, in line with the method proposed by Dienes (2021), we are performing a very specific series of analyses with H1 priors being modelled on the effect sizes previously observed by Hazen et al. (2009). As such, replacing these Bayesian t-tests with a mixed effects model would not be appropriate, and would result in an analysis that would no longer be comparable to that conducted by the original authors. Bayes factors will be calculated using an online Bayes factor calculator (URL: https://harry-tattan-birch.shinyapps.io/bayes-factor-calculator/), as now clarified in the manuscript on P. 24. Finally, regarding the suggestion to report the original Frequentist analyses reported by Hazen et al. (2009). All of our inferences will be based on rigorously calculated Bayes Factors following the guidance provided by Dienes (2021), but for each analysis where we report a Bayes Factor, we will also report the corresponding t and p-values. However, after careful consideration we propose not to report additional tests (such as the original

frequentist analyses) from which no inferences will follow, as to do so would be to invite non-pre-registered inferences that would be inappropriate in a registered report.

29) It would be helpful for the authors to explicitly state whether certain choices are in line with the study by Haazen et al. For example, it is unclear whether the decision to schedule trainings twice a week is in line with the study by Haazen et al.

Author Response: As stated in the start of the Methods section (see P. 14), all methodological decisions are in line with Hazen et al.'s (2009) study. Further clarification has been added regarding the decision to schedule training sessions twice a week (see P. 21).

30) Please add a reference for the Bayesian analyses on p.11

Author Response: Thank you for mentioning this – we have added a reference (end of first paragraph on P. 14) clarifying the specific Bayesian analysis protocol we will be employing.

31) What is the PSWQ >60 score based on? Please include a reference.

Author Response: Thank you for pointing out the need for clarity here – a justification for this cut-off has been added as a footnote on P. 15.

32) Is the maximum of N=200 based on previous studies?

Author Response: No, this limit of N = 200 has been enforced for practical reasons, this acknowledgement of practical limitations being the reason for which Schönbrodt and Wagenmakers (2018) introduced the maximal n *design. However, based on the Bayesian estimate of N analysis reported on P. 16, we believe there is a reasonable likelihood of obtaining sensitive evidence before this number of participants is reached.*

33) Some small spelling/interpunction errors were found. The authors should check the text again for these errors. For example, on p.3 in the Wittchen et al. reference and on p. 7 ('disorder' instead of 'disorders').

Author Response: Thank for pointing these out – the errors specified have been corrected, and the manuscript has been closely reviewed and further errors fixed.