# Peer review of "Minimal mindfulness of the world as an active control for a full mindfulness of mental states intervention: A Registered Report and Pilot study"

Thank you for the opportunity to review this manuscript. I have structured my review according to the PCI RR criteria, as follows.

## 1A. The scientific validity of the research question(s)

The authors conflate two different research questions: (1) to “test the centrality of metacognition in mindfulness practice“ and (2) to “test for the first time the effects of mindfulness against a true active control (if it should turn out to be one)”. I think that these two questions should not attempt to be answered by the same study. The authors cleverly suggest an intervention that minimises metacognition in order to test the role of metacognition in mindfulness practice. I think this is neat, but one should refrain from using this new intervention as a control for the effects of mindfulness (question 2) until question 1 is answered because otherwise one does not know what is being controlled for. The analogy would be to validate a questionnaire and use it to measure a construct in the same step. First, a questionnaire is validated, and then used to measure a construct.

## 1B. The logic, rationale, and plausibility of the proposed hypotheses (where a submission proposes hypotheses)

The hypotheses proposed for the main study look good to me. However, I think that the hypotheses and conclusions drawn from the pilot study were inadequate. A pilot study by definition does not have the power and methodological soundness to test hypotheses similar to the main study. Conclusions should not be drawn based on p values or significance. Rather, directions of effects and effects sizes should be considered for helping to design the main study. The lack of power in the pilot study may well be the main reason why its results were so confusing: most of the hypotheses were not confirmed, including the manipulation checks, and some effects even went in the opposite direction. I think that the pilot study in fact provides little information, the most important being showing the feasibility of the design, and the need for larger samples and better instruments.

For this reason, I think that the pilot study is given too much room and importance in this manuscript, which looks more like a report on the pilot study than a plan for a future study. I suggest that the authors tone down the conclusions of the pilot study (eg “If the Mindfulness of the World condition was not shown to have superior performance compared to Waitlist, this would indicate primarily functions is as a placebo mindfulness condition” is an overstatement in my opinion, because the lack of evidence for superior performance could have simply been due to lack of power). I also suggest they re-structure the manuscript so that it mainly describes a future study, which is informed by results from a pilot study. The details of the pilot could go in an appendix.

On a related note, in the pilot results the authors state that “A sensitive difference in the direction of the null hypothesis (H0) of no effect in the hypothesised direction was recorded between World and Waitlist condition for the PWB Environmental Mastery subscale, whilst evidence for an effect in the hypothesised direction between the Mental States group compared to the World group (i.e., a greater increase in the former)”, but the hypothesised direction was the opposite: “Conversely, the Mental States intervention should not increase the ‘Environmental Mastery’ subscale compared to the World intervention, given that subscale reflects an awareness of the world.”

## 1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable)

The methodology looks feasible, the pilot has been useful to this end.

The baseline-observation carried forward method is widely regarded as a suboptimal method for handling missing data (eg <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4785044/>). I suggest the authors consider using linear mixed models or multiple imputation techniques.

“If a survey is not completed within two days, a reminder email is sent, and after 3 days the survey is closed and participants are no longer allowed to take part in the study.” For an optimal ITT analysis, this needs to be changed so that participants can still complete questionnaire outcome data (and are encouraged to do so) even if they abandon the intervention itself. Also, it would be very informative to know why they abandon the course. This can be asked in a follow up email.

How will p values be calculated? This needs to be pre-specified. I am not familiar with Bayesian statistics so I cannot comment on their soundness.

## 1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses

The randomisation procedure needs more detail, in particular how the authors will ensure true randomness, and the concealment of the randomisation sequence for allocation.

I am not familiar with Bayesian statistics so I cannot comment on their degree of detail.

## 1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).

The modification of the Toronto mindfulness scale would require piloting especially if testing a pre-registered hypothesis (ie not exploratory). Otherwise authors run the risk of inconsistent results, as it happened in the pilot study. Or perhaps a new scale needs to be developed and validated to measure what they intend.

I would also pilot the modifications to the Observed sub-scale, although this is an exploratory analysis so it is less crucial.

In order to avoid ceiling effects, I would still include PSS in the larger study because it is more sensitive to non-clinical degrees of psychological distress (GAD and PHQ measure the clinical range better).