Evaluation of spatial learning and wayfinding in a complex maze using immersive virtual reality. A registered report.

PCI-RR Review Response - Round 3

Responses to the Editor and Reviewers are color coded as follows:

- Editor Comment (EC#) and Reviewer Comment (R#C#) are **gray**
- Responses to Editor Comment (ER#) and Reviewer Comments (R#R#) in **black**

Editor

EC1: Thank you for revising you Stage 1 RR to address previously identified issues of power. The manuscript has been seen again by both reviewers, who are happy with the changes made (GB reiterates the point that your chosen level of 0.8 power with alpha .05 will limit the set of possible publication venues, but I assume you have already considered this point).

ER1: We have reconsidered our position after Reviewer 1's suggestion and changed our thresholds to alpha = 0.02 and power = 0.90 (Cortex's proposed thresholds), which results in a sample size of 62. We believe the workload is manageable despite the increase in sample size. Thanks for your suggestion.

However, we are still unable to provide IPA for the manuscript, for the reasons sketched below:

EC2: 1) The power analysis relates to performance of the VRT, but this covers only one of the hypotheses. The second (set of) hypotheses relating to cyber-sickness and sense of presence ask quite different questions, but no consideration is given to the effect sizes of interest for these comparisons, or the adequacy of the sample size to these questions.

2) When considering this issue, you need to bear in mind the following further complications: (i) The tests are predicated on being able to detect a significant difference (in either direction) OR equivalence (by TOST), and your sensitivity for each of these possible outcomes may differ. (ii) Your conclusions will depend upon the combination of outcomes across multiple tests, specifically you state that the tests will be conducted on the 3 subscales of SSQ and the 4 subscales of ITC-SOPI. You need to make it explicit how your conclusions will be informed by the combination of outcomes across the subscales, and to correct the required alpha for multiple comparisons if appropriate. (iii) It is not clear where the PTT fits in to your research questions/hypotheses tests.

(If these complications cause too many problems, then you could consider relegating these further questions to a secondary exploratory status, and removing from the Stage 1 plan.)

ER2: After further discussion on this matter, we have decided to relegate these questions to our exploratory part of the study. PTT analysis is included as part of the exploratory analysis. We have modified/added our section Method → Analytic strategy

EC3: 3) The within-subjects design improves the power of your study, in principle, but it does create other issues. Specifically, is there a possibility of transfer effects between days; that is, might learning the task on one day (in one format) be expected to influence baseline performance (and thus opportunity for learning) on the second day? Unless I missed it, you do not even specify whether the same or different maze will be presented on each day, but these details seem potentially very important.

ER3: The Editor is right, as this isn`t clearly described in the Procedure. Participants will solve the same maze (start location order may differ as it will be pseudorandomized). To avoid transfer we thought of leaving 3 months between evaluations (when stated, VR studies usually leave up to a month between sessions, most only a week to solve this issue). We also chose 3 months to give us some time since, operationally, we will be able to do only a few examinations per week if this RR is approved to start in September (also due to the updated sample size). However we are confident transfer effects will be absent or minimal after this time, since a) it is a quite big and difficult maze to solve and b) in some of the original Wamsley studies, measured learning was similar when 2 evaluations were done within the same day. We have slightly modified the Method → Procedure to better reflect this.

EC4: 4) As a very minor issue, you state that participants will be randomly allocated to one of two task orders. I assume that allocation is not truly random if you intend to ensure that there are equal numbers for each order. Therefore, the allocation schedule may need to be stated more precisely.

ER4: To attend this issue we chose to perform a stratified permutation block randomization (https://clinicalresearch-apps.shinyapps.io/rrapp/) with two strata (for gender analyses). This is stated in the last paragraph of Method → Participants, which was edited for better understanding.

Reviewer 1

R1C1: Thank you for the changes you have made to the protocol - I think the within subjects design seems like a more sensible approach in this case. My one further suggestion would be merely an advisory one (perhaps the editor can weigh in) about the power calculation (and thus sample size) - the eventual outlet (https://rr.peercommunityin.org/PCIRegisteredReports/about/pci_rr_friendly_journals) may well have more stringent requirements than the 'bare minimum' alpha =.05 and power = 0.80 calcualted in the revision (e.g., Cortex uses alpha =.02 and power > 0.90). If this is a possible concern, then I'd recommend re-running the power calculation with 0.9 power, which seems approrpriate given the hypotheses.

R1R1: Upon your comment we have reconsidered our original alpha and power proposition and instead chose to set our thresholds at alpha = 0.02 and power = 0.90, which results in a sample size of 62. We believe the workload is manageable despite the increase in sample size. Thanks for your suggestion.

Reviewer 2

The authors have accurately addressed my concerns from before.

The comments addressed by the authors from Round 2 have also in a way answered some of my other concerns.

I believe this will make for a useful and important manuscript for the VR/Spatial Cognition community. I have some final notes:

R2C1: I now understand why the authors have used three trials. Considering this is based on previous research, it makes sense for this manuscript. However, I would strongly encourage any conclusions drawn about spatial learning to address this limitation.

R2C1: Yes. As with the original VMT, the task's design will be mentioned as a limitation to this study.

The introduction addressed my concerns clearly and also has a nice flow.

Thank you for providing OSF and GitHub links within the manuscript.

R2C2: Gender differences are an important one, and I appreciate you including it. However, I also appreciate that this was not part of the proposed hypotheses. It would be interesting to see if they are having an impact as some Desktop software can eliminate the classic water maze gender effect.

R2C2: Yes, it will be interesting to verify whether this effect persists in VR.

R2C3: I would like to praise the authors for the inclusion and construction of a Spanish version of the PTT. This is great, and I hope it will provide you with some interesting perspectives on variable spatial ability within different virtual environments (Does better PTT ability facilitate better desktop or iVR performance?). I understand this meant changing how the experiment would be run, so thank you for this.

R2C3: Thanks, and we agree it will add an interesting new angle from where to look our results from.

R2C4: I look forward to reading the completed manuscript.

R2C4: Thank you for your kind and very constructive comments.