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

PCI-RR Review Response - Round 4

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: 1) You have now relegated the statistical tests relating to cybersickness, sense of presence and perspective-taking to exploratory status, and removed these analyses from the design table. However, you still have a paragraph in Methods (p19) that describes the planned analyses, including the alpha level, approach to multiple comparisons etc. (You also refer here to "within-group comparisons", which I am not sure I understand.) This is an uneasy half-way house, because you are designating these analyses as exploratory, and yet you are effectively pre-registering them, but only in a relatively imprecise way (and without a priori consideration of their sensitivity). It would be preferable to preserve a clean distinction between registered and exploratory components by not descibing these analyses a priori.

I understand that in order for the reader to appreciate why the extra measures are included, you may wish to refer briefly to exploratory analyses of perspective taking ability, self-reported cyber-sickness and sense of presence, but you could leave the precise form of these analyses open until Stage 2.

ER1: This was a misunderstanding on my part. I incorrectly assumed that relegating them in practice meant that I could only take away our hypotheses related to those tests (i.e. directionality of results) while still describing what statistical tests we could/would be doing. I will leave those details for Stage 2 (Changes in Method → Analytic Strategy).

EC2: 2) It seems obvious why you would wish to compare cybersickness and sense of presence betweeen task versions, but it is much less clear why you would compare perspective-taking, which seems more like a (presumably) stable measure of spatial ability. I can't find any clear rationale for the inclusion of this as a dependent variable (although you do also mention its possible inclusion as a predictor for a regression analysis, which makes some more sense). Nor do I follow why you introduce the PTT as a test of 'spatial learning skill' - it does not seem to include any measure of learning.

ER2: The Editor is right. This was a typo as the PTT doesn't measure learning. We meant to state we'd study a "spatial ability", namely perspective taking, but this might be confusing, so we've changed our phrasing in Materials → Perspective Taking Test (PTT) and elsewhere. Our rationale to include it was to just have a tool to assess spatial skill in our participants and see how that is related to the with the VMT, a suggestion made by Reviewer 1, which we agreed to include. It will be part of a regression analysis, and not as a dependent variable.

EC3: 3) Your power analysis for the equivalence test seems (as far as I can tell) to be appropriate, but the description is not coherent or detailed enough. You state that you determined the SEOSI as "the mean critical effect size (maximum effect size that would not be statistically significant) product of of previously mentioned comparisons, resulting in a SESOI of d = 0.47." Aside from the typo ("of of"), the "mean critical product of previously mentioned comparisons" is too hard to follow, and needs better explanation.

Similarly, when you describe the critical power analysis, you state "... we conducted a series of sensitivity power analyses based on the two one-sided tests procedure for equivalence testing (TOST, Lakens, 2017) for dependent samples. With a statistical power of 90% and an alpha set at 0.02, the estimated sample size is 62 pairs/participants." This description does not seem precise. For instance: What is "a series of sensitivity power analyses"? What do you mean by "sensitivity power" (is it a typo)? What do your analyses actually consist of? What is it that you have power to detect? Is it the power to detect equivalence within your equivalence bounds? If so, then what is your power to detect differences (since your tests will also include these)? How does your conclusion about equivalence depend upon the set of outcomes across your different dependent measures? Do they all need to be equivalent in order to conclude equivalence overall? etc...

These points are not well explained in the appendix, which merely refers the reader to a downloadable spreadsheet from Lakens that will allow them to recalculate power for themselves (once they can work out for themselves how to use the spreadsheet). In any case, it would not be acceptable to offload the explanation of the TOST procedure (and power for it) to an appendix, because this is a critical part of the study design.

ER3: We've created a new section (Method → Participants → Sample size calculation) were we explained thoroughly (and perhaps more precisely) how the calculation of the SESOI came to be, explaining what data from previous studies we extracted, how we calculated the critical effect size for each of them and how that resulted in our SESOI. We also edited Method → Analytic Strategy to clarify our intentions of running dependent samples equivalence testing using TOST on 3 dependent measures (we deleted Speed because it might be redundant and not informative as it is defined as Completion Time/Distance Traveled). We briefly explain what the TOST procedure is, how our SESOI sets our equivalence bounds and when we will consider a result equivalent and when not. We also added when we will consider both versions to be equivalent (when all dependent measures are equivalent). We understand the key importance of this point, so please, if we're still missing something, or our explanation is still unclear, let us know.

EC4: 4) When I try to formulate my recommendation text for your study, I find that I am somewhat at a loss to pinpoint exactly what the point of your study is. I understand that you are testing for equivalent spatial learning (within certain bounds) between desktop and iVR versions of the VMT, on the basis that if they are equivalent then the iVR version could be considered as a substitute for the desktop version in some experimental contexts. But what if the iVR version shows greater evidence of spatial learning than the dektop version? Or lesser evidence of spatial learning? Would these outcomes mean that it could not be used as a valid version of the VMT?

Overall, it is clear that your main purpose is to compare spatial learning between task versions, but it is less evident why, in that it is not clear what you will conclude about the appropriateness of the iVR version given each of the different possible outcomes (including non-equivalence).

ER4: In the case of non-equivalent results, if the iVR version shows greater performance in all or some of the dependent measures could mean that immersion (greater field of view, response to head movements, higher presence, etc.) makes this task more ecological and therefore, better reflects the act of navigating. When conceiving this study, this is what we originally thought was to be expected (as did one of our Reviewers), before reading on the existing literature, where results show that there are usually no differences between versions, and when there are, they're worse for more immersive solutions (as explained in the Introduction).

If worse performance metrics are observed for iVR, we could attribute it to several instances (technical, user VR experience, etc.); however, if this is the case we expect it to be associated with increased cybersickness symptomatology (as measured by the SSQ). Again, this association has been found in some studies where iVR task performance was worse (particularly with the nausea subscale). We have taken measures to prevent this (blurring, FOV occlusion, pre-conditioning for

those without gaming/VR experience) and made some testing, which hopefully will decrease the incidence of cybersickness (at least compared to the Desktop version).

We have not been very explicit about all possible outcomes and their meaning (just hints in the Introduction) and this is only briefly pointed out at Table 1 (we expanded a little bit more for this revision), as we thought this belonged to the Discussion once results were known. If you think it should be included, could you please indicate to us?