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

PCI-RR Review Response (Rounds 1 and 2)

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**
- Notes in **red**

Round 1

Editor

EC1: *Both reviewers clearly feel that the plan has promise as a potentially useful contribution to the literature, and that the hypotheses of your study are well stated and clear, although there are some queries over the details of your methods. Reviewer#1 has a number of constuctive suggestions to make, including the suggestion that gender be not only balanced but analysed as an additional variable of interest (or controlled as a covariate). I think that your response to this point should depend critically upon precisely what hypothesis you want to test and whether gender is a relevant consideration for that hypothesis.*

ER1: After consideration, we will examine the effect of gender and spatial ability on our results.

EC2: *Reviewer#2 has a number of stylistic comments to make, and I agree very much with this reviewer's impression that the multiple-framing of the Introduction (in terms of dream literature, and in terms of validation for new method) was quite unclear and potentially confusing. One might think that, if your purpose is to create a task more likely to be incorporated into dreaming then a critical part of its validation would include an assessment of the rates at which it is incorporated into dreaming; but the methods make it clear that this is not part of your purpose. Perhaps try to be more clear about your aims, and more linear in your introductory narrative to establish these aims.*

ER2: The first section of the Introduction regarding navigation has been reduced as exploring its underpinnings is not an integral part of this study and might deviate the framing we originally intended. We have also made changes in the introduction so that it better reflects our aims by mentioning a) the advantages of using iVR over 2D media (ecological validity, skill transfer) despite the fact that there seems to be no difference in performance between the two and b) and how iVR is able to produce higher emotional arousal which might help increase task incorporation while dreaming. While it's not currently our purpose to explore rates of incorporation, it remains an important motivation for our work, since we believe that if we manage to validate our iVR-VMT, given the iVR properties previously exposed, it might help us with incorporation rates in a future study (Abstract and Introduction).

EC3: *This reviewers' point 5 is also critically important from an RR point of view. Do we really believe that a meaningful 'equivalence' could be established by ruling out effects smaller than the very large target level? Would we really consider any differences between tasks that are below this level of effect size to be irrelevant? This seems somewhat unlikely. Perhaps rather than motivating your smallest effect size of interest from expectations based on prior literature, it would be more relevant to consider from first principles what size of difference you think would be of no practical consequence to know about if it exists.*

ER3: We agree that our SESOI calculation might not be the most straightforward, however, we believe that it is the most useful in our specific case. We have two reasons:

1) Our rationale is based on the fact that variability in the main effect of interest (completion time improvement) in previous studies is quite large, both within and between groups. This is why d>1.1 effect sizes are observed in these studies: only large differences in improvement can be statistically significant (this is further elaborated in Reviewer 2's response). Thus, observing effect sizes smaller than d=1.1 might fall within that variability and be of no practical consequence for our specific task and measure. To avoid selecting a lower value arbitrarily we followed the *small telescopes* approach, which returned a value of d=0.77.

2) The effect sizes used to estimate our SESOI come from a between-subject comparison in the original VMT studies, whose design we pretend to replicate in a future study (use the iVR-VMT to test in a 2x2 ANOVA how task incorporation influences performance in Wake and Sleep participants). Therefore, and in light with what was exposed previously, if both versions of the VMT are equivalent, our iVR version should be able to detect the same between-subject differences/effect sizes.

Hopefully our intent is made clearer now. We understand the issues with this approach, which might not be ideal. We believe that it is possible that a difference leading to an effect size of d<0.77 might be of no practical consequence (taking into account the wide standard deviations in previous studies), but the value could as well be lower. If the Editor knows a better way to establish a realistic SESOI, we would be happy to follow the recommendations. However, if the value is too low it could become difficult to manage logistically (using a d=0.5 already requires a sample size of 69 per group, more than 4 times what was originally planned).

NOTE: Response to this matter has been updated in Round 2 (see below)

EC4: *(Related to this is Reviewer#2's point 3, which asks why the tests are configured as tests of equivalence, when there would seem to be a priori reasons to expect that the iVR version might be superior.)*

ER4: As written in our response to Reviewer 2, evidence from studies comparing task performance between Desktop and iVR versions of the same task tend to point to no significant difference between them, and when there are, they suggest poorer performance in the iVR modality. This might be dependent on the specific task and, more importantly, to the presence or absence of cybersickness symptoms. We do expect a higher sense of presence in the iVR condition, however, that might not be enough to improve performance in this task.

EC5*: In passing, I noted a few linguistic oddities in the Abstract, which you may wish to amend (there may be more similar oddities in the main paper):*

"commonground" >> "commonplace"

"understudied" >> "under-studied" or "little studied"

"One of such mazes" >> "One such maze"

"stimulant and engaging simulation" >> "stimulating and engaging experience" ?

ER5: We have edited the manuscript and corrected these and other expressions.

Reviewer 1

R1C1*: The authors present a pre-registration report to examine the learning ability and the usability of the Virtual Maze Task used originally by Wamsley et al. (2010). They will compare the Desktop version of the task to a more immersive VR version (using a HMD). The authors present an important*

research question that is often overlooked when using virtual tasks, do people actually learn better with greater immersion? The authors outline their research questions and hypothesis well, demonstrating two clear and concise hypotheses that can be easily tested following the reading of the methodology. The protocol is well described, though I have some minor comments about this (see below for section breakdown). The sample size is calculation is efficient but again, there are some minor comments not about its calculation, but more so its demographic. I think the clear presentation of method and hypotheses would provide a reader with confidence that no additional analysis will be explored, as only what the authors propose (assessment of learning in VMT and questionnaire use) will actually answer their research question. Both are commonplace across the literature as a methodology for assessment.

R1R1: We have attempted to solve these issues with our answers below.

***R1C2**: Statistical analysis is well outlined and in my opinion, valid for the experiment proposed. It is straightforward and easily replicable from the description. Though perhaps a different approach (repeated-measures examining individual trials and how they vary) may be important here, and may actually reveal more about the task - as spatial "learning" may not actually occur until familiarity with the environment etc. increases. This is particularly important when there is no recall or retest (probe) trial being used. This is common in the human and animal literature. A single score of "improvement" may not be enough data to warrant a true behavioural measure of participants actually "learning" the space and goal location, particularly not after three trials. Though perhaps I have misunderstood the procedure.*

R1R2: There is a retest (session 2) for the trials in the pretest (session 1) in order to test for task improvement. Although a repeated-measures 2x2 ANOVA (factor 1: Time (pretest, retest); factor 2: Group (Desktop, iVR)) could be performed on each of the raw values of our VMT variables, we chose to do a 2 sample t-test on the improvement values (retest minus pretest) following what previous studies have done and also because we are not currently interested in exploring the interactions product of the previously mentioned ANOVA. We have changed the manuscript so that this calculation is clearer (Method → Virtual Maze Task → VMT variables).

NOTE: Our experimental design was updated to a within-subjects design to better accommodate the reviewers and editor suggestions. For more information, see Round 2 below.

Specific comments:

Introduction

***R1C3**: Authors mention the lack of ecological validity of desktop tasks, but do not evaluate the greater ecological validity that may stem from iVR tasks. There are a lot of iVR studies out there, particularly some that also use iVR and real locomotion (e.g. Delaux et al., 2021). I think some mention of these and the above point would help your research question.*

R1R3: We have expanded our Introduction mentioning the greater ecological validity that comes with the use of VR tasks, specifically regarding navigation studies (Introduction).

Methods

***R1C4**: Authors mention they will achieve an equal number of male and female participants. However, I think gender should be recorded and analysed as an additional variable or control variable/co-variate. This is a well known effect in the human literature, even in virtual mazes (Mueller et al., 2008; Woolley et al., 2010). It has also been shown that it is specifically spatial learning that gender can have an impact on and not memory (Piber et al., 2018). I think this is an important aspect to examine or control for for both hypotheses.*

R1R4: We agree with the reviewer in this regard. We will include gender both as a covariate (to account for its impact on the dependent variable) and as an independent variable (to examine its effect on our results). (Method → Virtual Maze Task → VMT variables).

R1C5: Is three trials really sufficient to demonstrate "learning" without a memory trial? Many more trials are always used in the literature. I can also see the view that learning is clearly occuring - but I think this needs to be better justified (again, unless I misunderstood).

R1R5: We chose to validate this task using 3 trials per session replicating how the assessment has been performed in previous studies. Most of them observed within and/or between-session “learning” or improvement with only 3 trials per session. We believe more trials could better characterize this learning; however, this would make the task considerably longer and deviate considerably from the original task.

R1C6: The authors do a really good job at controlling for and assessing cybersickness.

R1R6: Thanks for your comment.

R1C7: For replication purposes, are both of the tasks used available Open Access? If not authors should recommend alternatives

R1R7: If the reviewer refers to the VMT tasks (Desktop and iVR), yes, the Unity project files are available in a Github repository (https://github.com/negatoscope/VRMaze) and an independent executable file was added, as suggested by another reviewer, in OSF (https://osf.io/2g6b8).

R1C8: Groups are controlled for cybersickness and well-assessed for this DV. Why are groups not controlled or assessed for spatial learning ability (which is a cognitive skill we know differs amongst individuals: Coutrot et al., 2018). Perhaps the introduction of a spatial task (Perspective Taking Task for Adults (Frick et al., 2014)) or even just a cognitive assessment (Trail Making Test etc). Just something that I think is important.

R1R8: Yes, we also believe that spatial learning ability will play a role in VMT performance, but hadn’t contemplated it as a covariate. We will use the Perspective Taking Test in its computerized version (Friedman et al. 2020), as it is openly available and shows similar performance to the Perspective Taking Task for Adults (Brucato et al. 2022) and correlates well with performance in virtual navigation tasks (Rekers et al. 2023). We made a Shiny App with the Spanish translated version of the PTT, available for preview (https://negatoscope.shinyapps.io/PTT_SOT_spanish/) along with its code in Github. We adapted the manuscript to reflect this addition (Method → Materials → Perspective Taking Test (PTT) and Method → Analytic Strategy)

On 26/06/2023 (just before this first round of review was sent) I received an answer from Andrea Frick, creator of the Perspective Taking Test for Children (and Adults), after asking her for the materials. In her answer, she mentions that she can share them, but doing so would imply her coaching me on how to run the task, which is “a lot of work”. If the reviewer thinks we should instead use this task I can reply asking for her help.

R1C9: I would be wary about the upper age-limit of your participants, over 18 may include older adults who may not be as familiar with the task. Either including them and controlling across groups, or setting a limit would be advised, as age has an impact on usability (Commins et al., 2020) and also spatial learning performance (see Figure 2E in Coutrot et al., 2018 for global data).

R1R9: Yes, here we made a mistake as we intended to put a upper limit on 30 years, following what has been done in previous studies. We adapted the manuscript to reflect this adjustment (Method → Participants)

***R1C10**: Nevertheless, this is a good and clean design to assess a very important question using a repeatedly used but rarely validated virtual spatial task from the literature. It is also an important, unanswered question which could eventually facilitate a framework for other researchers testing the reliability of their virtual tasks. It may also perhaps, save research teams time by implementing whatever version of the task fits into their setup and budget, if they have no real impact on learning. I would recommend this go forward with perhaps some additional thought in the areas mentioned above.*

R1R10: Thanks for your review and suggestions.

Reviewer 2

This interesting and timely article seeks to resolve the uncertainty in the literature about whether a VR version of a virtual maze task (VMT) compared favourably with a classic desktop version of the task. It is well-written throughout, and I appreciate the open materials narrative which is far too often overlooked in registered reports. As a disclaimer, I know little about the VMT literature, but have some expertise in VR and cognitive psychology in general. Some specific comments below.

Thanks for your interesting comments and suggestions.

***R2C1**: The manuscript as it stands is written in a slightly awkward mixture of past and future tense – I presume this is to make the edits to the final version after data collection a bit easier, but it didn't feel like there was much consistency in how these tenses where used throughout. I don't have strong feelings about whether this requires changing, but the editor may.*

R2R1: Indeed, this was done as a guide for future edits (and also followed what other RRs I checked have done). Edits have been done to the manuscript to improve consistency and readability.

***R2C2**: I found the framing of the introduction confusing – the narrative around the VMT is compelling, but I found the link to dreams as the main motivator for using VR rather tenuous. Indeed, the framing of the manuscript sits awkwardly between a paper seeking to resolve a dispute in the literature and a validation of a new method. I'm not sure it succeeds with either narrative, but do not know the extant literature well enough to say which is a more appropriate goal of the paper.*

R2R2: We agree that the multi-framing in our introduction might be confusing and unclear when establishing our rationale for this study. Ultimately, our main objective is to support validity evidence of the iVR version of the VMT, regardless of the possible absence of changes in performance, in order to get a more ecologically valid tool in the context of human navigation. Furthermore, we think that several of the iVR properties (such as sensory stimulation, and higher emotional arousal) give us a chance to improve task incorporation rates while dreaming in future studies. We argue that the possibility of creating a more stimulant, emotion inducing experience such as those that use iVR, might help people to dream about it. We have changed the manuscript, in hope that it better reflects this argument, mentioning (1) the advantages of using iVR over 2D media (ecological validity, skill transfer) despite the fact that there seems to be no difference in performance between the two, (2) and how iVR is able to produce higher emotional arousal which might help increase task incorporation while dreaming, and (3) by cutting the spatial learning/navigation framing, since our current aim is not to meticulously assess navigational skills in our participants (Abstract and Introduction).

***R2C3**: RQ1 and H1 are obvious enough (although surprising to me that the authors would be predicting equivalent performance – I'd have assumed that iVR would outperform desktop due to immersion)*

R2R3: We initially thought this as well, but according to the spatial learning/navigation literature that compares Desktop vs iVR versions of the same task/game/experience using modern HMDs, most articles found no significant difference in performance between the two, or if there are, they lean towards better performance the Desktop version. This is possibly due to the higher rates of cybersickness in participants in the iVR conditions (as acknowledged by some of the authors) or the specific task that is being tested. We've updated the manuscript to better reflect this statement (Introduction).

R2C4: I was unable to open the build linked in the manuscript – this probably reflects my difficulties with unity hub rather than the application, but it would be worth uploading 'fixed' .exe builds for desktop and quest that will be used in the manuscript to streamline this process for more casual users

R2R4: I have uploaded a new version of the Unity project in Github (https://github.com/negatoscope/VRMaze) along with the .exe build (and required files) in a .rar file available at OSF (https://osf.io/2g6b8). We also updated the manuscript (Method → Materials → Virtual Maze Task)

R2C5: My biggest issue with this article is from the power calculation and associated sample size. The authors base the sample size calculation. The Wamsley papers from which the d=1.1 estimate is drawn from is a correlational analysis bears no resemblance to the methods of the current study. I understand the challenges of predicting an effect size, but would recommend the authors use an effect size derived from other studies comparing VR to desktop environments OR the smallest relevant example of how VMT performance can vary from one condition to another in a between-group design. As it stands, the current sample size seems far too low to provide anything like a resolution of the issues in the literature or a validation of VR in this context. As it stands, the TOST would miss any different that was just below a 'classic' large effect size of 0.8, which feels like very shaky grounds to declare equivalence. This is a pretty simple paradigm and I see no reason not to be more conservative in this regard.

R2R5: I agree with the reviewer that this is a contemptuous aspect of this study for which we contemplated a couple of options before deciding. As the reviewer suggests, we first examined the studies that compare both versions (Desktop vs iVR) in terms of task performance, excluding variables or measures that evaluated skill transfer, presence or user experience. As mentioned before, most studies reported no significant differences between versions, and those that did, leaned towards better performance in the Desktop version, presumably due to higher rates/intensity of cybersickness in participants in the iVR condition. This led us to decide to use equivalence testing and to test for an absence of effect.

When we attempted to determine our SESOI (for the TOST's lower and upper bounds) that would allow us to calculate our sample size, we looked for studies with tasks that were somewhat similar to ours (in that they explored navigation/wayfinding and used completion time as an estimate). However, in some we didn't find enough data to calculate the effect size (n, mean and/or standard deviation, a bar plot, etc.) and in those that we did, we found large effect sizes after accounting for motion sickness (d=1.2-1.7, Carbonell-Carrera et al., 2021; Clemenson et al., 2020).

For this reason we chose to use effect sizes from VMT performance in previous studies (as also suggested by you). The two effect sizes used for reference (d=2.2, d=1.1) come from a pre-post performance VMT comparison in a between-group design (Wake vs Sleep) where those that incorporated or had mentations about the task had better performance than those who didn't (not a correlational analysis). It's important to note that we will follow this design in a posterior study. Despite using a large effect size of d=0.77, we believe that it remains useful for our purposes since VMT performance scores (completion time, distance) are wildly variable within and between participants (they show very large standard deviations in Wamsley et al. studies), so only when the difference is

large enough we should expect to find significant effect, whether between Desktop and iVR or between Wake vs Sleep groups.

To demonstrate this, if we take as reference the results that led to the effect sizes from the studies by Wamsley et al. mentioned in the last paragraph, the minimal effect sizes that are also statistically significant ($p<0.05$) are $d=1.05$ for Wamsley et al. 2010 and $d=1.04$ for (Wamsley et al. 2019). This suggests that values under $d=1$ might be of no practical consequence for this specific task and measures. To avoid selecting a lower value arbitrarily we followed the *small telescopes* approach, which returned a value of $d=0.77$.

Hopefully this rationale is clear, and we will adapt our manuscript accordingly. If it is not, or if the reviewer has a better alternative, we would be happy to follow the required adaptations.

NOTE: Response to this matter has been updated in Round 2 (see below). In short, we have changed our experimental design to a within-subjects design and selected another parameter of VMT performance (from previous studies) that led us to define a narrower SESOI of d=0,47 (for more details, see below).

Round 2

Editor

EC1: Thank you for sending this revised Stage 1 manuscript, with replies to reviewers. Having looked at your replies, I see one potentially major issue, and I think it most sensible to return the manuscript directly to you for further consideration before asking for reviewers to devote more time to this.

At the first round, Reviewer#2 raised the critical issue that it seems relatively unimportant to test for equivalence, where this is defined by an effect size for the difference smaller than d = .77, given that this actually describes a very large effect size. That is, your test would be prepared to declare equivalence between tests even in a statistically large difference between the tests existed - it is hard to see this as a useful form of equivalence,

In your response, your main line of reasoning is that d=0.77 is appropriate for your purposes because VMT performance scores are "wildly variable within and between participants... so only when the difference is large enough we should expect to find significant effect". This comment seems to reflect a misapprehension of what your measure of effect size (d) represents. Cohen's d is a standardised measure of effect size that is expressed in units of SD (it is the mean difference between groups divided by the pooled standard deviation). Therefore, d of .77 remains a very large effect size, regardless of how variable the performance is between participants. (If the performance is more variable between-subjects, this just means that the mean difference that d of .77 represents is proportionally larger.)

ER1: We agree with this comment. We understand that this is a standardized measure, and that d=0.77 is nevertheless a large effect size and independent of the standard deviation value. What we meant when referring to the large between-subjects variability was that, if we assume SDs were to remain as variable, by setting smaller equivalence bounds we would also approach a lower significant (and equivalent) mean difference, which could go beyond what we think is relevant, given the mentioned variability, at the expense of a significantly higher sample size. However, we now understand that, as is, this method could lead to inaccurate results.

EC2: Given this fact, as far as I can see, your response does not address the problem at all, and you remain in a position of having an equivalence test that could rule out only very large (seriously non-trivial) differences between tasks. You suggest that you may be at the edges of practicality of

sample sizes required for testing smaller effect sizes than this, but that might simply indicate that you are not in a position to run a meaningfully useful study of the sort that you would like. (You might also want to think about whether the VMT task itself is worth trying to adapt to iVR if performance is, as you say, so wildly variable within and between participants.)

ER2: We did not intend to give the impression that we were actively trying to not run a meaningful study. This is our first attempt at a Registered Report and we honestly believed that a large effect size could be justified as a useful SESOI, given the task's circumstances and our aim (see if two almost identical tasks give out the same measures). We now understand this could come in detriment of general accuracy and also the message that we would be able to transmit. Regarding whether it is worth adapting the VMT to iVR or not, we believe it is, because it is this manipulation that we think could improve incorporation rates when we do a conceptual replication of the Wamsley set of studies.

EC3: Alternatively, you may be able to improve statistical sensitivity to smaller effects by designing a within-subjects study? And/or perhaps you could consider running your study with lower power e.g. .8), although this would reduce the strength of conclusions you could draw (and also affect the range of possible destination journals). In any case I am sending this back to you for reconsideration of this key point. Perhaps the following article from Zoltan Dienes could be useful in helping you think about how to define realistic effect sizes that might be worth ruling out (the article takes a Bayesian approach, but the guidance for defining meaningful effect sizes of interest applies equally for a frequentist approach): https://doi.org/10.1177/2515245919876960

ER3: After your suggestions and paper recommendation, we did some consultation and came to the conclusion that while deviating from the original studies procedure, doing a within-subjects study would help us control at least the intrasubject variability, and thus, improve our study's accuracy. We have also searched for other VMT variables that could help us to set a lower SESOI than before, while still using existing data to avoid setting our equivalence bounds (EB) subjectively. In the end, we found that a SESOI of d=0,47 setting the EB for a dependent-measures TOST significantly improves the accuracy and interpretability when compared to our last proposal. The rationale that got us to the SESOI of d=0,47 is as follows:

We used the variable "baseline performance" which is the completion time at the last (third) training trial of the VMT. We chose this measure since it is also used to define our "improvement" measures (Improvement = Baseline performance - Mean of the 3 test trials) while also reflecting the VMT variability independently of groups and conditions. Since baseline performance values weren't readily available in the papers, I had to extract them from their plots using WebPlotDigitizer. In the end, we got values from 6 papers using the VMT.

Table 1. Mean, standard deviation (SD) and sample size (n) from samples used to estimate the SESOI

	Mean	**SD**	**n**
Wamsley 2010a	277,37	150	99
Wamsley 2010b	263	169,24	48
Nguyen 2013	249,9	156,6	30
Wamsley 2016	230,21	171,49	97
Murphy 2018	219,29	163,7	27
Wamsley 2019	317,82	152	17
Pooled values	**255,45**	**161,36**	**318**

For each “baseline performance” variable in each paper we extracted its mean, standard deviation and sample size. We then computed t-test comparisons of the “baseline performance” variable between all samples (15 in total), computing their t-value, p-value and effect sizes. We then calculated their critical test statistic value (CTV) and critical effect size (CES), which represents the maximum effect size that is not statistically significant between groups (Lakens 2018). Finally, we calculated the mean for these 15 CES values and used it as our SESOI for our TOST power analysis and equivalence test. This way we get an informed value that describes how variable VMT performance can be, while also helping us establish reasonable boundaries for equivalent testing.

Table 2. T-test comparison between samples, critical statistic and effect size calculations. SE: standard error; CTV: critical test statistic value; CES: critical effect size.

Comparison		**Mean diff**	**SE mean**	**t**	**p**	**d**	**CTV**	**CES**
Wamsley 2010a vs.	Wamsley 2010b	14,37	27,52	0,5221	0,6024	0,092	1,985	0,349
	Nguyen 2013	27,47	31,58	0,8698	0,386	0,187	2,007	0,418
	Wamsley 2016	47,16	23	2,0504	0,0417	0,293	2,016	0,288
	Murphy 2018	58,08	33,21	1,7487	0,0828	0,38	1,998	0,434
	Wamsley 2019	-40,45	39,45	-1,0252	0,3074	0,269	1,993	0,523
Wamsley 2010b vs.	Nguyen 2013	13,1	38,29	0,3421	0,7332	0,08	2,011	0,468
	Wamsley 2016	32,79	30,13	1,0882	0,2784	0,192	1,991	0,351
	Murphy 2018	43,71	40,24	1,0861	0,281	0,261	2,013	0,484
	Wamsley 2019	-54,82	46,57	-1,177	0,2436	0,332	2,018	0,57
Nguyen 2013 vs.	Wamsley 2016	19,69	35,12	0,5605	0,5761	0,117	1,99	0,416
	Murphy 2018	30,61	42,44	0,7212	0,4738	0,191	2,024	0,537
	Wamsley 2019	-67,92	47,04	-1,4436	0,1558	0,438	2,021	0,614
Wamsley 2016 vs.	Murphy 2018	10,92	36,96	0,2955	0,7681	0,064	2,008	0,437
	Wamsley 2019	-87,61	44,39	-1,9735	0,0509	0,519	2	0,526
Murphy 2018 vs.	Wamsley 2019	-98,53	49,33	-1,9972	0,0523	0,618	2,027	0,628
							Mean CES	**0,470**

A TOST power analysis for dependent measures reveals that for an alpha=0.05, power=0.8 and equivalence boundaries of ±0,47, a sample of 39 pairs is required (calculations can be replicated using the TOSTER R package or this TOSTER Excel spreadsheet https://osf.io/qzjaj).

We have changed our manuscript to reflect changes in the sample size calculation and changes in the experimental procedure (adding an additional session for the within-subjects comparisons), and added our detailed rationale in Appendix 1 in the Supplementary Material.

Thank you very much for your comments and suggestions. We have attempted to solve our study’s main issues (effect size, within-subjects designs) based on your comments. We hope that this current version satisfies the editor's (and reviewer’s) requirements, and consider the manuscript for further review. We look forward to your comments, and we are totally open to new suggestions that may continue to improve this study.