Thank you for submitting your Stage 1 manuscript to PCI-RR. As the recommender assigned to this manuscript, it is my role to perform an initial triage assessment to determine whether the submission is ready to be sent for external review. This assessment is primarily with respect to the RR aspects of the proposal, rather than its specific topical content. On the basis of this initial assessment, I would say that the manuscript is generally very well prepared, but that there are a number of specific points that you should consider before it sent for external review. These all relate to the analysis plan/design table, which is what I focused my assessment on.

1. It is great that you have included manipulation checks, but you must be explicit, in each case, about how your conclusions would be affected if the check were failed. Normally, a manipulation check would be to confirm an effect that would be expected if the task is working as intended, or is otherwise a necessary precondition for your experiment to be deemed capable of testing the experimental hypotheses of interest. If your manipulation check has this status, then it should be made explicit (your first manipulation check seems like it would probably have this status, but I am less sure for 2-4). If it does not have this status, then it is not an important manipulation check, and should probably be omitted.

Response: Thanks, we have removed (what was previously) questions 2, 3 and 4, as we agree that these checks are not critical to the experiment. We have expanded the checks on the real-world version of the task to also include the sensorimotor prediction effect (generally gripping large or denser-looking objects with more force than smaller or less dense-looking objects) and to check that people can articulate an expectation that larger (or denser) objects are likely to be heavier prior to touching the objects.

2. Manipulation check 1 has two parts that support two different conclusions (about the SWI and MWI respectively). It should therefore probably be split into two (perhaps labelled 1a and 1b). In general, a separate hypothesis should be defined for each test that supports a separate conclusion.

Response: We have renumbered the questions as ‘a’ and ‘b’ for the SWI and MWI throughout, as suggested.

3. As noted, you should consider whether manipulation checks 2-4 are necessary. Is it essential that there are no differences here (and would any differences invalidate or qualify the conclusions you can draw from your main experimental hypothesis tests)? If so, then why is it considered adequate to have 90% power only to detect a rather large effect size (dz = 0.66). If it is indeed necessary to confirm no differences then you should frame an equivalence test, rather than just failing to reject the null (or use a Bayesian approach). Also, since the verbal statement concerns ‘general differences’ between real and VR, it seems like you should have a simple independent t-test (which tests the overall difference between conditions), rather than an ANOVA (which would be sensitive to any pattern of differences between means).

Response: As suggested we have removed these tests from the table of questions.

4. Although manipulation checks are not your experimental hypotheses, it would still be conventional to label them as hypotheses (e.g. H1). You can still contextualise as a manipulation check, explaining how your experiment is affected if the hypothesis is not confirmed.

Response: Thanks, all labelled sequentially now.

5. The targeted effect sizes for your experimental hypotheses reflect mean estimates from prior research. Unless these prior studies were registered reports, then it is likely that these are over-estimates of the true effect (indeed, the very fact that you are choosing to follow up these findings may also mean they are likely to be over-estimates). It might be preferable to target a lower-bound or otherwise conservative effect size. In any case, you need to provide a rationale for why the targeted effect size is appropriate (i.e. that failing to detect an effect of this size would be an important message for the field).

Response: We have added a more general rationale for the types of effects we consider to be of interest and a more detailed rationale for the sample size.

6. H2a, H2b: A conventionally 'medium' correlation is selected as the smallest effect size of interest, but no rationale is provided for why this is a theoretically or practically relevant SEOSI.

Response: As this was an exploratory question and not one of the primary questions addressing our primary question we have removed it from the table, as suggested below.

7. For some hypotheses, the associated theoretical conclusion will be drawn if the associated test is significant for either SWI or MWI. This 'disjunction' logic (X if either Y or Z) probably requires alpha adjustment (e.g. Rubin, 2021. https://doi.org/10.1007/s11229-021-03276-4).

Response: following discussions about whether it would be appropriate to treat the two tasks as disjunctive, we have chosen to approach them as individual hypotheses, as described in the Rubin paper. We have included words to this effect in the table of questions.

8. Exploratory research questions should not be included in the Stage 1 plan, but can be added at Stage 2. This would also help to simplify the current Stage 1 manuscript.

Response: Thanks we have removed these, but kept a table of them on the OSF page and will run them at stage 2, as suggested.

9. In the Data treatment section of the main text, it is stated that "Data will be checked for extreme deviations from normality based on skewness and kurtosis scores. Assuming data adhere to these assumptions the tests outlined in the table of questions will be run. Non-parametric alternatives will be used if data deviate substantially from normality." You need to provide precise statements of what will constitute sufficient

deviations for you to switch analysis strategy, and you also need to state what the non-parametric alternative will be. Note that the above assumptions are not necessary for your regression analyses (which assumes normality of residuals).

Response: Thanks we have added cut-off values for skewness and kurtosis and specified the non-parametric alternatives.

10. In the Data treatment section of the main text, it is also stated that "Bayes factors using a symmetric Cauchy prior will also be used to quantifying the strength of evidence for the alternative and null hypotheses." These do not feature in the design table, so I assume they play not role in your inferential logic. It is OK to mention that these BFs will be calculated, if it is your plan, but you should make it clear that your main conclusions will be driven only by the outcomes of the tests specified in the design table. If the purpose is to try to quantify evidence for null results, and you regard this as important for your conclusion, then perhaps a Bayesian approach would be more appropriate as your main strategy (or you could include some frequentist tests of equivalence).

Response: Thanks, we have added some text to the methods to specify that these are not driving our conclusions, but that they are being used as an additional tool to help quantify the strength of evidence.