Editor:

Thank you for submitting this revised Stage 1 plan, which has now received external review from two relevant experts. I think that their reviews are very helpful, both at a theoretical and practical level, and that consideration of these comments should help you clarify the Stage 1 plan.

To my reading (although I could be wrong), the reviewers make very similar points regarding the framing of your hypotheses, which mean that - at least with the present design - your experiment is only able to inform about the *relative* influence of priors vs sensory information. You should clarify what conclusions your hypothesis tests can and cannot support, and perhaps re-name the hypotheses if this will help to avoid confusion.

Also, Reviewer#1 was somewhat confused by the status of the exploratory questions, which seems to have resulted from your removing these from the manuscript (as suggested at the previous round), but keeping the detailed design table as supplementary material. I understand the desire to keep these design components online, but it will be much clearer to pre-register only the primary hypothesis tests, and to save exploratory investigations until Stage 2 (keeping the design table for exploratory parts as supplemental material occupies an uneasy half-way house).

I hope that the reviewers' thoughtful comments will be helpful for you in fine-tuning your plans. If you choose to address these comments in a revision, then you should include a responses document that specifies how you have addressed each comment.

Review by anonymous reviewer 1, 12 Mar 2024 13:05

The plan is for an interesting study comparing reaching behaviour in real vs virtual environments. A number of analyses are carefully planned and described. I am just unsure of the rationale for designating most of these as “exploratory” despite the high level of detail provided – see below.

14-43. VR vs real environments. Another notable issue is that VR settings vary in the nature and quality of online visual feedback. People may either not be able to see the hand they are reaching with, or it may be tracked and represented inaccurately - this affects studies in which online control of movements is included / a potential issue.

Response: Thanks, we have added a note on this.

44-75. Opposite predictions for priors. The two possibilities of (1) reduced use of priors and (2) reduced sensory info are not mutually exclusive – they could both occur. Really, you are asking about the final balance – does the balance in potential changes in these two shift people towards using the prior more, or using the prior less? It makes sense that these two outcomes could come about because of various combinations of changes in the prior and/or the sensory info (likelihood). However, it’s over-simplistic to suggest that either only one changes (reduces), or only the other does.

I would instead suggest to hypothesis either a shift towards using the prior less (expected if prior is weakened more than sensory info is – they could both be weakened somewhat) or using the prior more (if sensory info is weakened more than prior is)

Response: Thanks, yes we fully agree that we can only talk about the relative balance of priors versus sensory information. We can't truly determine if one has weakened or the other strengthened. This was our original intention and we have talked about the relative balance of priors and sensory information in some places but realise that we were not consistent in doing this all the way through. We have made some adjustments (e.g., lines 60, 63, 77-79) to the wording that should now make it clear that we are focused on the balance rather than the absolute values.

129-142. Exploratory analyses:
Mindful of the suggestions
at https://rr.peercommunityin.org/help/guide_for_reviewers#h_7586915642301613635089357

"*Have the authors minimised all discussion of post hoc exploratory analyses, apart from those that must be explained to justify specific design features? Maintaining this clear distinction at Stage 1 can prevent exploratory analyses at Stage 2 being inadvertently presented as pre-planned*."

As far as I see, the details and justification for the exploratory analyses at lines 129-142 might be needed in order to justify the design (i.e. collecting fingertip force data). Is there a way to explain more clearly why this info is being provided now (if not to 'pre-register' these analyses)? (e.g. in order to explain why the measures are being collected?)

Response: In our initial submission we had included these questions related to learning rates and transfer as pre-registered analyses. However, as they were secondary questions and not underpinned by the power analyses, the editor suggested that they did not need to be pre-registered and could be conducted later as exploratory analyses. Reflecting on the guidance that you cite, we see that we did not need to discuss these ideas in the introduction (or refer to them in the methods) so have removed all the text about these questions. We have just left in the presence questionnaire in the methods and noted that its purpose was for any exploratory analyses.

The fingertip force data is primarily being collected to address the main research question about sensorimotor prediction (H_{1B}) so these additional analyses are not needed to justify the collection of the force data.

151 Primary and exploratory questions and Table (Table of questions_revision.pdf).

The mapping between the document and the table is very unclear. On the one hand, the first item in the document refers us to hypotheses 4 and 5, skipping over 1-3. On the other hand, 1-3 from the table are not mentioned in the text.

I would suggest for the table to follow the order that is in the document, and everything in the table to be referred to (at least in summary) also in the document.

Response: Hypotheses 1-3 in the table are just manipulation checks, which we thought weren't worth discussing in the manuscript but were worth recording in the table. So that that the numbering matches up we have simply re-ordered the hypotheses in the table so that the manipulation checks are 3-5 (even though they will be run first, which we have noted in the table). And we now make a brief mention of the manipulation checks in the main document.

More fundamentally - I will preface this by saying that I don't have experience with this format. However: the decision to call many of these analyses exploratory, but at the same time to carefully list them in the table in terms of hypotheses, measures, conclusions to be drawn, does not make sense to me. The advice above notes that it is good to "*prevent exploratory analyses at Stage 2 being inadvertently presented as pre-planned*". There is so much pre-planning here that I struggle to see what is exploratory about these analyses? The only distinction that calling them exploratory seems to me to provide is that, depending on the outcomes, the authors might not feel that they need to report them all (e.g. if some are unclear or uninteresting). I would suggest that these are all carefully pre-planned analyses, and you could plan to report them all, regardless of the outcomes. Or, to maintain these as exploratory, there would be much less need to pre-register these plans for them.

Response: Thanks, yes in response to the previous comment we have removed all mention of these additional analyses and will include them as clearly marked exploratory tests in the final manuscript if we decide that they are useful.

180 Methods. These seem to be sound and well grounded in previous studies, including those by the authors.

Response: Thank you.

235 I am concerned that grasping via control of a white sphere vs normal hand could lead to large differences in visuomotor control. This is a valid part of how VR interactions can be different to real ones, but should be discussed more among the reasons for real-VR differences (see comment in intro). I also wonder if the authors have considered an open-loop situation instead to better match the tasks on this dimensions (but perhaps applicability of this to these tasks/illusions is unclear/unknown).

Response: We appreciate this concern. This protocol was based on our previously-published work using a white sphere to guide the hand to the object in a whole-hand grasp (Buckingham, G. (2019). Examining the size–weight illusion with visuo-haptic conflict in immersive virtual reality. Quarterly Journal of Experimental Psychology, 72(9), 2168-2175. https://doi.org/10.1177/1747021819835808). In pilot testing we verified that after a few practice reaches, participants comfortably able to perform the precision grasp of the force transducer handle without sight of a realistic hand. Almost

certainly the kinematics of the reach will be affected by the hand visualization, but as all the variables we are measuring occur after the object has been contacted we think this should have minimal impact on our conclusions. This will, however, be an important point for the discussion. We did consider an open-loop design, and experimented with methods of visualizing the hand in VR, but these introduced different issues related to reliability of visual information. Furthermore, the critical question we aim to examine is about the differences between high fidelity virtual environments and the real-world, and so wanted to keep the integrity of both without other manipulations of visual feedback.

It took me a while to find a clear statement of the numbers of trials to be collected (there is in one place a mention in passing on line 267 of "10 lifts" as part of a calculation). Later at line 340 there is the plan for 30 test trials (10 per object). Some justification that this number is likely to have enough power to answer the hypotheses of interest (especially the main ones)?

Response: The grip force variable is calculated from just the first lifts of the smaller and larger (or more/less dense looking) objects so trial numbers are irrelevant for this variable. For the heaviness ratings, the choice of 10 lifts followed previous studies that have used 8 or 10 lifts per object, e.g.:

Arthur, T., Vine, S., Brosnan, M., & Buckingham, G. (2020). Predictive sensorimotor control in autism. *Brain*, 143(10), 3151-3163.

Buckingham, G. (2019). Examining the size–weight illusion with visuo-haptic conflict in immersive virtual reality. *Quarterly Journal of Experimental Psychology*, *72*(9), 2168-2175.

Naylor, C. E., Proulx, M. J., & Buckingham, G. (2022). Using immersive virtual reality to examine how visual and tactile cues drive the material-weight illusion. *Attention, Perception, & Psychophysics*, *84*(2), 509-518.

350 Perhaps I should know this, but why choose 3.29 standard deviations from the mean as the outlier criterion (corresponds to xx% - 99% maybe?) A criterion based on quartiles / IQR can deal better with non-normal data. On the other hand, if this has been commonly used in similar studies, it sounds OK.

Response: Yes this was a cut-off that has been used in previous similar studies examing fingertip force scaling in the context of these kind of stimuli (e.g., Arthur et al. 2020). We chose it as a conservative marker for outlying data points so that only very extreme values will be considered outliers - 99.7% of normally distributed data will fall within that range. We have added detail to the manuscript on this point.

Arthur, T., Vine, S., Brosnan, M., & Buckingham, G. (2020). Predictive sensorimotor control in autism. *Brain*, 143(10), 3151-3163.

Review by anonymous reviewer 2, 01 Mar 2024 21:14

The submission proposes an experiment examining differences in use of prior expectations about object weight and material properties across real-world and VR object lifting, using the size-weight and size-material illusions to probe 'weight' given to prior expectations. The proposed study is technically accomplished, and the analysis pipeline is clearly specified and well designed. And I find the central question—how much do we 'trust' prior expectations about the world in VR?—to be deeply interesting. I think there are some fundamental problems with the theoretical conceptualization of the study and resultant hypotheses, however, that mean it cannot do the job the authors intend, and so I focus on those.

The study is formulated in the framework of Bayesian inference, where the 'weight' given to the prior (here, prior expectations about object weight etc.) should reflect the relative reliabilities of prior expectations vs. sensory input. The two hypotheses are then described as alternative propositions about the prior (low-precision vs. high-precision). But really they aren't. Instead, they speak to the different terms in Bayes' rule. LPP hypothesizes that the *prior* will be less reliable (or treated as less reliable) when the subject knows they are in a virtual world, whereas HPP supposes that the *sensory input* (the likelihood, in Bayesian inference) is less reliable in VR. These are both reasonable propositions, but they are orthogonal. Both can simultaneously be true (or false, or any combination thereof). So the question appears to be ill-posed.

In more formal terms, there is (conceptually) one dependent measure (weight given to prior expectations) but two unknowns (i. reliability of sensory input; ii. reliability of prior expectations). If we allow that both sensory and prior reliabilities can change across VR and real-world—and I think we have to—it's not possible to infer what caused any measured change in reliance on prior expectations (which the proposed theoretical interpretation depends on). Consider the case where sensory reliability and prior reliability are both reduced in VR (which seems plausible). Depending on the exact, quantitative nature of those reductions, this could result in prior expectations receiving more, less, or the same weight (though, note, here 'no-reweighting' would not be because nothing changed). Running the process in reverse, finding that people relied more heavily or less heavily on prior expectations does not allow us to infer how the underlying reliabilities of likelihood and prior have changed (except perhaps at the extreme ends of the possible outcomes).

A key step here would be to measure the reliability of sensory input empirically across the different situations, so it is no longer an uncontrolled variable. If the reliability of visual size information, for instance, was matched across real-world and VR, changes in the weight given to prior expectations in different contexts—which I'd argue is the deeper question here—could be ascribed unambiguously. I must say I'm not sure how to do this for the reliability with which material properties (granite etc.) are specified, or even if that's necessary, but I think it's worth thinking about.

Response: Thank you for this comment. The other reviewer also identified the same issue. Our original conceptualisation of this study was that we were interested in the *relative* influence of prior expectations on perception and action. As you clearly outline, it could be that both are changing at the same time and we cannot know if a shift in the

influence of prior expectations is due to changes in the prior or the sensory input. So our intention was only ever to consider the relative balance between the two. But we acknowledge that we didn't make this clear in the manuscript.

Regarding your suggestion about measuring the reliability of sensory input, we agree this would be ideal, but our experimental protocol is not well suited for this endeavour. The visual size is perfectly matched between the two conditions (the 3d model displayed in VR is the same 3d model printed for the task in the physical environment), so there should be no difference in that regard. But really the issue is the *perceived* reliability of the information – even if the visual information was all perfectly matched (and we could measure that) some participants could up/down weight it if they decided that the VR version was less reliable (even if its not). So we think that we can only gauge the relative influence of prior versus sensory and have reworded aspects of the introduction to hopefully make this clearer (e.g., lines 60, 63, 77-79).

Relatedly, and taking a step back, I'm a bit troubled by the VR vs. real-world manipulation as conceived here. It seems to presume VR to be 'monolithic', but visual information could be less reliable, same, or (in-principle at least) more reliable in VR compared to the real world, depending on the exact properties of the system used, the scene parameters and content etc. So in my view it isn't meaningful to think in terms of general conclusions about VR vs. the real world. They need to be qualified by the type of understanding of the constituent signals and their reliabilities in a given situation, as above.

Response: Yes we agree that we can't make overly general conclusions about VR as a single entity when there is so much potential for variation. Our study design was aimed at providing two very closely matched environments (visual and haptic information) so that any differences were most likely related to the virtual nature of the VR task rather than any substantive difference in the information available. We have added some text to acknowledge this and will make sure this is key part of the discussion (line 24, 82-87).