Reply to Recommender:

Thank you for taking the time to look over our revision and thank you for your comments, I have addressed each below:

I appreciate you have now added specific hypotheses for the pairwise posthoc comparisons. However, as these are specific hypotheses (and Bonferroni corrected for multiple testing) these require sufficient statistical power in their own right. You cannot simply refer to the power of the omnibus test here. We already discussed this in earlier rounds of the review. It is crucial that the preregistered hypothesis matches the statistical test used.

To address this issue, you have two choices: The more thorough approach would be to add a specific power analysis for your posthoc tests. A simpler alternative is to remove the posthoc tests again and simply specify a basic omnibus hypothesis. The posthoc comparisons can be included at Stage 2 as exploratory analyses, provided they are explicitly labelled as such, and they should not be mentioned in Stage 1 at all. However, keep in mind that this change makes your preregistered hypotheses relative vague:

Hypothesis 1: Is there a significant difference in subjective illusion experience between the experiment conditions?

Hypothesis 2: Are there significant changes in the somatosensory response when comparing different conditions healthy participants? **(This is how H2 currently reads already)**

Thank you for pointing out this omission, we have now added power analyses for all post-hoc comparisons which has been updated in text and in the design planner. These power analyses have made an adjustment to our sample size which has also been updated:

Sample size update:

"Overall, based on the power analyses in section 2.5, a total sample size of <u>34</u> participants will be tested. This sample size adheres to the higher end of sample size estimates (Hypothesis 2 (2.5.2) showing <u>34</u> participants needed <u>for post hoc tests 2a and 2b</u>)."

Update to Hypothesis 2 power analysis:

"This is the first study to investigate illusory finger stretching using SSEPs, so appropriate effect size estimates are not available. We therefore conducted power calculations based on a smallest effect size of interest, in line with the recommendation of Lakens (2014). Here, we have chosen an effect size of d = 0.5 (a medium effect, see Cohen, 1988), since this is the smallest effect size we are interested in detecting, which we have converted to a Cohen's f of 0.25 <u>for Hypothesis 2's power analysis, and have maintained at 0.5 for the subsequent post hoc power analyses</u>.

*<u>Hypothesis 2</u>: A priori power analysis using G*Power shows that for a repeated measures, within factors one way ANOVA, with an effect size (f) of 0.25, alpha of 0.05, power at 90%, and 1 group with four measurements, a total sample size of 30 participants is needed.*

*<u>Hypotheses 2a and 2b: A priori power analysis using G*Power shows that for a two-tailed difference between 2 means (pairwise) t test, with an effect size of dz = .5, alpha of 0.05, power at 80%, a total sample size of 34 participants is needed.</u>*

*Hypothesis 2c: A priori power analysis using G*Power shows that for a one-tailed difference between 2 means (pairwise) t test, with an effect size of dz = .5, alpha of 0.05, power at 80%, a total sample size of 27 participants is needed.”*

These changes are matched in the design planner.

- **Please fix the effect size for Hypothesis 1 in the Design Table to match the plan in the manuscript.The former still states f = 0.73.**

This has been corrected.

- **Where you describe the illusion index you write in lines 339-342:**

The normalised (baseline corrected) data will be used for analyses, with a new scale from -100 to +100 with 100 indicating strongly agree, 50 indicating a neutral opinion, and scores below 0 indicating strongly disagree with the statements on the questionnaire.

Wouldn't the neutral point be 0 in this case or am I misunderstanding how this is calculated?

If, for example, a participant gives an averaged median score of 80 for the illusion questions, an averaged median score of 20 for the disownership questions and an averaged median score of 10 for the control questions, then the control score of 10 would be subtracted from the illusion score of 80 to give 70, and from the disownership score of 20 to give 10. Initially the score of 80 indicated agreement, the score of 20 indicated disagreement. With these new corrected scores, the new score of 70 would still indicate agreement and the new score of 10 would still indicate disagreement. If we set the neutral point from 50 to 0, then the disownership new score of 10 would no longer indicate disagreement, it would indicate agreement as it is above the new neutral cut off 0. Therefore, the score of 50 will be retained to indicate a neutral opinion of the question.