Reply to the Recommender's / Reviewers' Comments

We would like to thank the recommender for their comments, and have addressed each point in italics below:

1. Main hypothesis (Hypothesis 2):

"[...]there will be a significant main effect of condition between pain patients and healthy participants, measured via SSSEPs, when comparing (2a) multisensory visuo-tactile illusory resizing to a non-illusion condition, (2b) unimodal visual illusory resizing to a non-illusion condition, and when comparing (2c) multisensory visuotactile illusory resizing to unimodal visual illusory resizing."

I found it difficult to ascertain what exactly this hypothesis actually is, both in terms of the general question and the specific statistical test used. You describe this as a "main effect" but the statement reads like an interaction: you want to compare the "effect of condition" (i.e. chronic pain vs control) "when comparing" the experimental conditions. Put more simply, this asks whether the illusion strengths vary between groups. Judging by the description later on this is not what you are actually testing?

To remove uncertainty around what Hypothesis 2 is, Hypothesis 2 has now been clarified, and predicts a significant difference between illusory conditions within the same participant groups as dependent t tests rather than main effects in an ANOVA, as can be seen below:

"There will be a significant difference in SSEP response when comparing (2a) multisensory visuotactile illusory resizing to non-illusion, and when comparing (2b) unimodal visual illusory resizing to non-illusion in the Healthy Group. There will also be a significant difference in SSEP response when comparing (2c) multisensory visuotactile illusory resizing to non-illusion, and when comparing (2d) unimodal visual illusory resizing to non-illusion in the Chronic Pain Group.

Analysis of EEG data will involve taking the amplitude of the 26Hz response for each subject and condition (coherently averaged across repetition), before running a dependent samples t test comparing MS to NI and one comparing UV to NI in the healthy group, along with a dependent samples t test comparing MS to NI and one comparing UV to NI in the chronic pain group. The dependant variable will be SSSEP amplitude in μV."

This change has been clarified throughout the report.

2. Clearly specify all your hypotheses

In general, it is advisable, especially in RRs, to reduce any complex interaction effects down to the critical 1-df contrast that can actually address the question. This enhances the sensitivity of your analysis plan and can help minimise the required sample size. The way I read your description throughout the manuscript, I believe you are indeed looking at main effects, which are in essence a 1-df t-test between experimental conditions *irrespective of pain group*. In that case, however, it is unclear why you even need a control group.

This issue is most notable when looking at the design table in the Appendix. The Analysis Plan column for Hypotheses 2a, 2b, and 2c are in fact all identical. You state here that given significant main effects (which are the same for all three hypotheses!) you will run Tukey's posthoc tests.

These Tukey posthoc tests are in fact the statistical contrasts your experiment should be powered to detect.

The complex interaction has been reduced to instead using t-tests as mentioned above, and therefore the Analysis Plan section of the design table is now not identical for Hypotheses 2a, 2b and 2c, instead these have been changed to reflect the dependent t tests to be run comparing each illusion condition as mentioned in the point above, in addition to an extra hypothesis (2d) for the chronic pain group. See an example of the change below:

"A dependent samples t test will be run comparing MS to NI in the healthy group. The dependant variable will be SSSEP amplitude in µV."

A similar point applies to Hypothesis 3 where you describe the hypothesis as *"[...]reduction of pain [...] before and after each illusion, when comparing multisensory, unimodal visual, and non-illusion conditions"*. This hypothesis and the associated statistical tests do not actually compare the three experimental conditions. This also makes the interpretation ambiguous because it is unclear what constitutes support for Hypothesis 3.

Hypothesis 3 has been clarified to compare pain scores for both types of illusion, as can be seen below:

"(3) We expect to find a subjective reduction in pain, measured via a 21-point numeric rating scale, comparing before and after scores for multisensory and unimodal-visual conditions.

Pain data will also be analysed using JASP (JASP Team, 2022). Since the data will be ordinal, non-parametric Wilcoxon signed rank tests will be used to compare mean pain scores before and after each condition.

If Hypothesis 3 is supported: Indicates that analgesia can arise from multisensory and unimodal visual illusory resizing.

If Hypothesis 3 is unsupported: Indicates that analgesia is either associated with one condition (either multisensory or unimodal-visual), or with neither condition."

These changes are reflected throughout the report.

3. Remove redundancy

While I commend you for specifying different comparisons and their potential outcomes, there is also a lot of redundancy in your text. I already mentioned the most striking redundancy in the design table in #2. But there are also numerous duplicates of statements in the text, especially in section 2.3.2. For most of the sub-hypotheses you describe it seems unnecessary to explicitly state how you interpret the outcomes because you are stating the obvious: if hypothesis 2a is supported, you found a significant difference. Note that such explicit statements can indeed be helpful, for example when you describe the positive control. If you fail to find a significant difference between the illusion and non-illusion conditions, this calls the results from Hypotheses 2 and 3 into question. However, for most of these comparisons I would argue that this repetition actually impedes the reader's understanding of what you're doing.

To reduce redundancy, replications of interpreted outcomes have been removed from section 2.3.2, and in place statements such as the following are included:

"Interpretations for hypothesis 3 can be found in the design table."

Interpretations for the positive control have been kept in section 2.3.2, as recommended.

4. Interpretation of potential outcomes in Hypothesis 3

There is also some inconsistency in how these interpretations are described for Hypothesis 3. You state that if Hypothesis 3 is unsupported you write *"Whereas if there are no changes in somatosensory cortex but pain reduction is seen. this shows that the driver of illusory induced analgesia is not coming from changes within the somatosensory cortex."* This sentence belongs in the previous paragraph because it means that Hypothesis 3 was in fact supported? (Minor side note: please also note the typo before "this shows" where the period should be a comma).

Inconsistencies for Hypothesis 3 have been removed and changes have been made to improve clarity, as can be seen below:

"(3) We expect to find a subjective reduction in pain, measured via a 21-point numeric rating scale, comparing before and after scores for multisensory and unimodal-visual conditions.

Pain data will also be analysed using JASP (JASP Team, 2022). Since the data will be ordinal, non-parametric Wilcoxon signed rank tests will be used to compare mean pain scores before and after each condition.

If Hypothesis 3 is supported: Indicates that analgesia can arise from multisensory and unimodal visual illusory resizing.

If Hypothesis 3 is unsupported: Indicates that analgesia is either associated with one condition (either multisensory or unimodal-visual), or with neither condition."

These changes are reflected throughout the report.

5. Power analysis

The power analysis is ostensibly based on previously reported minimal effect sizes, but many of these are simply the effect sizes you found (and in an as-yet unpublished study). This is not necessarily a problem but you need to justify why this is a useful *mininum effect size of interest*. Would finding a smaller effect than this mean that we should accept the null hypothesis?

Moreover, I am confused how you chose the effect sizes used to power Hypothesis 2. In the first paragraph of section 2.4.2 you list the effect sizes from your previous work. The smallest of these is Cohen's f=0.27. In the following sentence you then explain that you chose larger effect sizes (f=0.42, 0.63, and 0.4, respectively) for your power analysis because they *"adhere to the lower end of the effect size range."* This seems to be a contradiction.

Effect sizes for Hypothesis 2 have been adjusted to reflect the smallest effect sizes of interest (Lakens, 2014) of d = 0.5 (a medium effect, see Cohen, 1988). The section has been updated with the amended analysis (dependent samples t tests in place of ANOVA) and can be seen below:

"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 (Lakens, 2014) of d = 0.5 (a medium effect, see Cohen, 1988).

*A priori power analysis using G*Power shows that for a matched pairs one-sided t test, with an effect size of d = .5, alpha of 0.02, power at 90%, a total sample size of 47 participants is needed for each participant group."*

There is also again a lot of redundancy in this section. Given that all those hypotheses described are using the same statistical test, you only need to run a power analysis based on the smallest effect size of interest (which should probably be lower than what you found in previous work). So instead of restating the power analysis for each effect size you only need to do this once.

On that note, it is also not strictly necessary to state the actual power achieved, if your analysis is based on 90% power. Either say you calculated the sample size needed for 90% power, or state the power you have at that sample size.

Redundancy has been removed and the remaining power analyses state that they concern the smallest effect size of interest, and do not mention the estimated achieved power, but state that power is at 90%, as can be seen in an example below:

*"A priori power analysis using G*Power for the smallest effect size of interest (f = .73) shows that for a repeated measures, within factors ANOVA, with an effect size (f) of 0.73, alpha of 0.02, power at 90% and 2 groups with three measurements, 8 participants are needed for each group."*

Finally, the sample size you settled on is *"85 participants (42 per group)"*. That sounds mathematically awkward.

The sample size has been adjusted based on the changes to the power analysis for hypothesis 2, and now reads as below:

"Overall, based on the power analyses in section 2.4, a total sample size of 94 participants (47 Healthy, 47 Chronic Pain) will be recruited, to adhere to the higher end of sample size estimates (Hypothesis 2 (2.4.2))."

6. Time line

The timeline is useful (although you may want to remove this at Stage 2, so perhaps this is better kept in the Appendix). However, this seems to contain an error: you expect that you will need 3.5 months for recruitment. This corresponds to the longest period in your Gantt chart which seems to be *data collection*. The *recruitment* period is only a bit over a month, presumably the 7 weeks you mentioned needing previously to recruit 50 participants. I understand that recruitment and data collection go hand in hand but as it stands this doesn't seem to add up.

*Timeline has been moved to the appendix and the confusion over recruitment has been clarified. The issue before was that the phrasing "which is based on data acquisition completed by this research group for a previous study also using EEG and resizing illusions, in which c.50 participants were **recruited** in 7 weeks" was inaccurate, this now correctly reads "which is based on data acquisition completed by this research group for a previous study also using EEG and resizing illusions, in which c.50 participants were **tested** in 7 weeks", with corresponding changes later in the text.*

Moreover, you aim to recruit 90 participants but your power analysis is based on n=85. Aiming for more is certainly wise, but please clarify what happens if you end up with a sample of 85 or more but less than 90. The sampling plan in the classical frequentist framework should specify the exact sample size you plan to collect. Some flexibility in this can be mitigated statistically (e.g. using Bayes Factors instead of frequentist tests) but if so this needs to be part of the experimental plan.

To provide clarity, the exact sample size is now mentioned in a new section, please see below:

"2.5 Sample Size

"Overall, based on the power analyses in section 2.4, a total sample size of 94 participants (47 Healthy, 47 Chronic Pain) will be recruited, to adhere to the higher end of sample size estimates (Hypothesis 2 (2.4.2))."

Clarity for what constitutes a final sample size is mentioned at the end of section 2.2 Experimental Procedure, "Data collection will be terminated when the full sample of participant have been recruited (47 Healthy, 47 Chronic Pain). If a participant completes <50% of the experiment their data will not be included, and additional participants will be recruited to fill any lost data."