## 1. Clarity of rationale/theory

- Predictive coding and central sensitization theory are described as accounts explaining the cortical misrepresentation of the body that might occur in chronic pain patients. I interpret this misrepresentation as a distorted homunculus (altered topographic map). Yet, based on the provided info, it is not clear to me how either of these accounts can explain such a cortical misrepresentation. Moreover, w.r.t. predictive coding, some descriptions seem a little unclear, such as what is meant by a mismatch between experienced and actual painful sensory inputs.
- I have difficulty appreciating the rationale underlying the proposed study. 1. Based on the provided info, it is not entirely clear to me what the "benefit" of inducing SSEPs is, as they seem to be primarily sensitive to temporal features of (multi)sensory stimulation (e.g. synchrony). 2. The intro repeatedly talks about a cortical misrepresentation of the body in chronic pain patients, that resizing illusions might affect this misrepresentation and that this might be related to a reduction in pain observed in these individuals. It is not clear to me how the proposed study can shed light onto these aspects (not least because the authors do not seem to be interested in group comparisons).
- Given that the intro does not explicitly state what experimental conditions will be included, the paragraph outlining the hypotheses at the end of the intro is not easy to understand. It also refers to "subjective embodiment theories". It is unclear to me what exactly is meant here and how these theories are related to the hypotheses.

## 2. Presentation of pilot data/accessibility and status of prior work

- In the intro (+ section on power analysis), pilot data are mentioned, which are not presented in 3.3 Pilot data. Similarly, prior submitted work (Hansford et al., 2022) that is mentioned in the intro and used to calculate effect sizes does not seem to be available as a preprint (at least no preprint including doi has been referenced), making it hard to evaluate these aspects. Moreover, Hansford et al. (2022) is cited as "in prep" in the text, but listed as "submitted" in the reference section, creating confusion about the status of this article.

## 3. Comprehensiveness/clarity/appropriateness/redundancy of inclusion/exclusion criteria

- Although mentioned nowhere, I think the authors intend to pull the finger of the right hand (?). "Chronic pain in (one of) the fingers of the right hand" thus seems to be an unspecified inclusion criterion.
- It is mentioned that data will be excluded if less than 50% of the experiment has been completed or more than 50% of the electrodes need removal. It is unclear what exactly this refers to (e.g. data set of a single participant and trials?) and why these criteria are sensible. For instance, if only 50% have been completed, the number of trials in the different experimental conditions per participant might be vastly different.
- It is mentioned that participants will be matched based on gender, but that the demographic survey will assess sex.
- In light of the interindividual differences reported in the intro, I do not understand why the authors intend to include chronic pain patients irrespective of chronic pain condition.
- Healthy and chronic pain participants share many inclusion/exclusion criteria, which are outlined twice (once for each group). This makes it a little hard to appreciate the differences.
- To transform the sampling characteristics into a "Participants" section, I think further points need to be covered, such as consent, ethics, and Declaration of Helsinki.

## 4. Clarity/appropriateness/detail of experimental procedure

***Questionnaires – handedness and pain***

- The Waterloo Handedness Questionnaire has not been described and it is unclear how exactly right-handedness (inclusion criterion) will be determined.
- It is unclear what participants will be asked when they have to indicate their pain score. Given that they will also be asked which finger is most painful, I would assume they have to rate the level of pain for this finger?

***Digit manipulation***

- It is stated that if multiple fingers are equally painful, the one that is easier to manipulate will be chosen and that more than one digit can be manipulated if needed. It is not clear to me why some fingers will be easier to manipulate or multiple fingers need to be manipulated and how the experimenter figures this out. I think this level of flexibility has the potential to create fundamental differences between some participants.

***Augmented reality system***

- The text refers to 2 white dots on the felt. In Figure 1, however, I see 4 white dots. It is also stated that the 2 dots will guide where the hand will be placed. How?
- When describing the augmented reality system, I think the camera needs to be mentioned too, so that it is 100% clear where the video/image comes from.

***Experimental conditions***

- The experimental conditions are hard to picture. Adding static visualizations/videos might be very helpful.
- The way the text (including design table) refers to the experimental conditions is confusing. In some places, it refers to multiple multisensory or unisensory illusory conditions, although there is just one of each. Similarly, there are 2 non-illusory control conditions and in several places the text refers to a single non-illusory control condition.
- For the multisensory illusory condition, the text states that the hand will be stretched and for the unisensory illusory condition that the finger will be stretched. For the non-illusion control condition with tactile input, it is not speci-

fied what happens to the visual input. Condition NI is described as a non-illusion control condition without tactile input, which seems somewhat incorrect as there will always be tactile input delivered via a tactile stimulator.

- It is stated that an indicative box tells the experimenter whether to pull the finger or apply no manipulation. This description seems incomplete, as the experimenter also needs to be informed when they have to touch the finger without pulling (condition NIT). This in turn means that in most cases the experimenter knows which condition will be presented. As such, it seems a little odd that the text states the experimenter will be "blinded".
- In the section on the experimental procedure, it is stated that the level of pain will be assessed before and after each illusory condition (in chronic pain patients), whereas the section on preprocessing steps refers to pain data that have been assessed for 3 conditions including one of the control conditions (MS, UV, and NI).

***Questionnaire – illusory experience***

- It is stated that at the end of the experiment, all conditions will be presented again in an ordered fashion. It is unclear what the exact order will be and why "ordered" is to be preferred over "randomized".
- The illusory experience questionnaire includes 6 questions, some of which refer to the finger and some of which to the whole hand. Shouldn't they all refer to the finger? Moreover, how will participants respond to the 6 questions? The section on planned analyses suggests a scale will be used.
- There are 2 control questions to assess compliance effects. Will they be used to remove data sets?

***Unaddressed details that seem necessary to ensure consistency and replicability***

- How long does a trial for a given condition last? How will the end of a trial be registered? How is it ensured that the temporal structure of different trials is the same and that actual pulling and visual stretching are synchronized? How long does it take until the video is augmented? What happens if the experimenter makes an error (e.g. forgets to pull)? What is the percentage of visual finger stretching? Is the augmented reality system an item that has been purchased (if so from where?) or is it self-built? What are the basic specs of the augmented reality system (e.g., size of screen; resolution of screen, size of felt, overall height, type/brand of camera, etc.)? Will all fingers be outstretched or just the one that is being pulled? What are the other parameters of the sine wave (e.g., amplitude)? Where will the experimenter be seated and how exactly will they pull a given digit? What software will be used to program the experiment? Who will operate the tablet? How will the pain rating be assessed (verbally/paper/tablet)? Will participants be seated in a shielded chamber? Will there be ocular/reference electrodes? Will participants' finger be continuously stimulated (without any breaks) throughout a given block and also when they have to complete the final illusory experience questionnaire? Will the stimulator be attached to the finger that will be pulled? Etc.

## 5. Appropriateness/clarity/detail of preprocessing steps

- The outlined preprocessing of the EEG data seems very minimal. For instance, I would assume the EEG data need to also undergo segmentation, artifact correction, averaging, and a Fourier transform (some of this is mentioned in passing in 3.3. Pilot data). Moreover, what software will be used/what are the electrodes of interest/what about lateralization? It seems critical to specify these things, not least because they (might) determine the number of tests being performed as part of hypothesis 2.
- W.r.t. the selection of EEG data, it is not clear to me what is meant by calculating "the top 5% of standard errors" and using this as a threshold (maybe the $95^{th}$ percentile?) and why it is sensible to use this threshold and keep a data set even if 50% of the electrodes have been removed.
- The outlined preprocessing of the pain and illusory experience data involves averaging. In the section on planned analyses, it is mentioned that Likert scales will be used for the illusory experience questionnaire (rendering the data ordinal) and that the pain data are ordinal. By calculating averages, these data are treated as "interval". Besides, it is unclear to me why it is sensible to first average the pain data, then group according to chronic pain condition, and then average again. This procedure might give a lot of weight to a few individuals with a certain pain condition and does not seem to fit in with the proposed tests for hypothesis 3.

## 6. Clarity of stated hypotheses

- Given that there are 2 non-illusion conditions, it is unclear what is meant by "non-illusion condition" when hypotheses 1+2 and associated analyses are outlined.
- Hypothesis 1 appears underspecified. Just like for hypothesis 2, shouldn't the subhypotheses for hypothesis 1 be specified separately for healthy and chronic pain participants, resulting in 4 instead of 2 subhypotheses?
- Hypothesis 3 seems underspecified. As far as I understand, it effectively consists of 2 subhypotheses (3a: reduction in pain [pre vs post] for the multisensory illusory condition; 3b: reduction in pain [pre vs post] for the unisensory illusory condition) instead of a single hypothesis.
- W.r.t. hypothesis 3, it is also unclear to me why the authors do not wish to include a control condition – I think that would facilitate the interpretation of results.

## 7. Appropriateness/clarity of planned analyses

- Hypothesis 1: It is not clear to me why a Friedman test should be performed given that all subhypotheses seem to relate to contrasts between specific experimental conditions. Similarly, it is unclear to me why the text states that 3 comparisons will be made. Which groups/conditions do these 3 comparisons involve?

- Hypothesis 1b: I think the authors might run into a double-dipping/regression to the mean problem (see e.g., Kriegeskorte et al., 2009; Stoll et al., 2022). This is because the illusory experience data from the unisensory illusion condition will be used for selection (subset of individuals experiencing the illusion in this condition) and selective analysis (comparison of illusory experience data in this condition to illusory experience data from a control condition for this subset). This renders this procedure circular and likely results in regression towards the mean (or variants thereof), i.e., statistical artifacts. One way to break this circularity would be to assess the unisensory condition twice in each individual, so that one data set can be used for selection and the other for selective analysis. However, due to the lack of detail on the scales assessing illusory experience, it is not clear to me why individuals need to be selected in the first place or why individuals with an average illusion score above 1 in the unimodal illusion condition should be selected. In any case, the selection of individuals seems to come with an uncertainty about the sample size – it seems unclear how the proposed study accounts for that.
- It would be good to always specify what software will be used, whether a test is one-sided or two-sided and how multiple comparisons will be dealt with, which I think is also relevant for power analyses.

## 8. Appropriateness/clarity of reported effect sizes and power analyses

- Hypothesis 1: I think it needs to be ensured that the ultimate contrast of interest is sufficiently powered (see also my comments on planned analyses).
- Hypothesis 2: A power analysis for a one-sided paired *t*-test has been performed. Hypothesis 2, however, is expressed in a way that seems to suggest a two-sided test.
- The descriptions related to the effect sizes for hypothesis 1 are not easy to understand because it seems unclear what they refer to. For instance, what does it mean that hand-based resizing illusions show a certain effect size? What has been compared here? Similarly, for hypothesis 3, it is not clear whether the same pain scale/multisensory resizing illusion have been used in prior work and what the effect size in Preston et al. (2020) amounts to.

## 9. Clarity of presented pilot data

- The text states that a pilot study has been conducted to determine the ideal frequency. However, as far as I understand, it has only been tested how well a frequency of 26 Hz works (?).
- It is not entirely clear to me what the exact data basis for the amplitude vs frequency graph is (currently Figure 3, although I think it should read Figure 2). For instance, have the data been averaged across all electrodes and all conditions and are they based on healthy participants? Have the data been cleaned?
- It is stated that for the pilot data, no illusion experience questions have been assessed. I thus wonder whether it is ensured that the illusory conditions still work properly when adding tactile input at a frequency of 26 Hz.

## 10. Consistent usage of abbreviations and capitalization

- Once introduced, it would be good to consistently use capitalizations/abbreviations, such as Healthy Group, OA, SSSEP, or MS (instead of e.g. going back to spelled-out versions or using variants of abbreviations, such as UVS or UV). For clarity, I think even common abbreviations (EEG) should be spelled out upon first usage.

## 11. Clarity/comprehensiveness of appendices

- Appendix A: The single figure here is labeled “Figure 2”. I think the figure numbering typically starts over in each appendix (Figure A1 in this case). The text here refers to 90 participants, but the main text to 94 participants.
- Appendix B: I think it would be good to give the design table a title and table number (i.e., Table B1) and use the table number in the main text. I wonder why the design table does not outline the rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis and the theory that could be shown wrong by the outcomes. I also think the statements about achieved power can be removed.