Major revision

Thank you for your well written paper. I have now received thorough comments from two expert reviewers. I will make further comments to do with the Registered Reports format in particular.

We thank the Recommender for his positive view of our submission and have taken his suggestions as well as those of the Reviewers into account to improve the manuscript. We believe that this has greatly improved our project.

1) *Make sure there is no analytic and inferential flexibility left; that is, anyone following your description would reach identical conclusions from your data. For example:*

- *p12 "the average amplitude of the signal measured at 2-5 neighboring frequencies to remove residual noise"*
  There is analytic flexibility here; decide precisely what you will do in advance.

We apologize for not clarifying this aspect in the protocol. Depending on the location of the electrode, there are up to 5 neighboring electrodes. Due to their location, some electrodes only have 2 neighboring electrodes. In the analysis, we will always choose the maximum number of neighboring electrodes for the removal of residual noise. This is now specified in section 2.7.1 on p.12 of the manuscript.

- *p13 "and the harmonic with the largest modulation (i.e.,Wilcoxon signed-rank test statistic)" What will you do if the test is non-significant? Why not just select the harmonic with the largest modulation?*

The statistical test is used to ensure that the selected amplitude is significantly different from zero at the frequency of stimulation. If the sustained periodic stimulation does not lead to an increase of EEG amplitude at the frequency of stimulation, we assume that the stimulation paradigm was not effective. Without this basic modulation, we cannot assess differences in the modulation following the oddball stimuli, and this is therefore our positive control for the EEG response. A purely visual assessment of the largest modulation could lead to misinterpretations.

- *In the Design Table, indicate exactly which test you are using to make inferences. In Row 1 you indicate a two factor model: Will the inference be based on just the interaction? On the interaction plus follow-up simple effects? Which simple effects? Here and elsewhere tie down your inferential chain *exactly*. Ask yourself for each row: Is someone else*

guaranteed to come to one single conclusion, the same as yourself, given the specifications here?

We thank the Recommender for pointing out the importance of ensuring that the inferential chain is fully registered. If we find a significant effect in the interaction, we will use a pairwise t-test to distinguish the direction of the effect for each condition (comparing baseline to oddball for high and low intensity oddball separately). This is now specified in line 1 of the hypothesis table as well as in section 2.8.1.

We further specified how the Wilcoxon signed-rank test is interpreted to decide whether there is a modulation of the EEG signal at the specified frequency, both in the hypotheses table and in section 2.8.2.

2) *Power should be calculated with respect to the effect you do not wish to miss out on detecting, in order to control Type II errors rates. If one calculates power with respect to an average past effect, it means you have not controlled for missing out on interesting effects less than this. For some guidance on thinking about this see:* https://doi.org/10.1525/collabra.28202

We fully understand the Recommender's concern that we could miss out on effects that are smaller than expected if we do not adjust our power calculation to the effect that we do not want to miss out on and we are grateful for the reference provided. To determine the sample size that would allow us to find the smallest effect we would be interested in, we used the method of calculating the 80% confidence interval of the observed effect size and using the lower bound estimate as the new estimated effect size for the power calculation (Perugini et al., 2014).

Based on table 1 in Perugini et al. (2014), which details the calculation of the sample size needed to detect the smallest effect size that would still be interesting based on observed effect size, sample size of the observed effect and the confidence interval (80%) for a targeted power of 0.8. According to this table, the initially planned sample size would suffice to find the smallest effect size we could be interested in for the detection of the peak at the frequency of the baseline stimulation using a Wilcoxon signed-rank test, comparing the amplitude at the baseline frequency of stimulation against 0.

To be able to detect the smallest still interesting effect size regarding the peak of the response at the oddball frequency, using our estimated intermediate effect size as the basis of the calculations, we would need to recruit over 160 subjects. Unfortunately, we do not have the resources (time / finances) to collect such a large sample size in our lab. Yet, we wish to point out that the intermediate effect size has been estimated based on an experiment that used an oddball

frequency-tagging approach in a different domain (visual, not pain), and it is therefore possible that the observed effect will be larger than expected. This expectation is based on the notion that a painful stimulus is regarded as a “threat” to the body, and will therefore recruit a larger response than e.g. a visual (non-threatening) stimulus since it has more relevance for homeostasis. Nevertheless, as a compromise, we suggest that we will not draw any definitive conclusions on the definitive absence of an effect based on a non-significant result in this statistical test.

Similarly, for the Linear Mixed model assessing the differences in perception ratings between oddball stimuli and conditions, over 2000 subjects would be needed to be able to detect the smallest effects we would still be interested in (based on our initial assumption of an intermediate effect size). As for the previously described statistical test, we will refrain from drawing any conclusion in case of a non-significant result. We therefore also removed the test as a “positive control” for our investigation since it did not seem reasonable to use it as such if we do not have a clear interpretation of a negative result.

All of these considerations are now added in the “Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis” section in the Supplementary Materials. Additionally, the hypotheses table was adapted to reflect these considerations appropriately.

Review by Björn Horing, 08 Oct 2023 08:16

The proposal titled "The effect of stimulus saliency on the modulation of pain-related ongoing neural oscillations: a Registered Report" by Leu, Forest, Legrain and Liberati lays out an experimental protocol to disentangle various aspects of the processing of noxious stimuli. These include stimulus intensity processing, stimulus saliency, and ultimately the perception of pain. The protocol uses behavioral readouts (continuous ratings) and electroencephalography (EEG) using frequency-tagging.

Previously, I have been reviewing the first iteration of the protocol; I now find that several of my concerns regarding the earlier draft have been adequately addressed. A number of issues remain, however (some old, some new). I have carefully worked through the manuscript and will list my comments individually or under each relevant stipulation listed at https://rr.peercommunityin.org/help/guide_for_reviewers#h_3015488595591613635204737 .

Note that there is no perfect protocol and I believe the authors' approach has merit regardless of whether all concerns can be fully addressed.

We thank the Reviewer for his positive in-depth review of our submitted manuscript and are grateful for the insightful comments and constructive suggestions. We have carefully considered your feedback and hope that our point-by-point response will address your concerns.

= = = = = = = = = = =

Individual major concerns

The authors suggest an oddball paradigm where a baseline heat stimulation occurs at a higher frequency (sinusoidal stimulation at 0.5Hz). At a lower frequency (0.125Hz), oddball stimuli are interspersed in two variants (conditions), namely higher-than-baseline intensity, or lower-than-baseline intensity. Core hypotheses concern the relationship of the oddballs to each other, with different interpretations depending on whether the higher oddball is accompanied by higher EEG/ratings than the lower, or not.

I have two concerns.

I. *The first concern is hard to address but should at least be discussed as limitation, or maybe the authors have an idea to actually solve it (e.g. two different suprathreshold/high oddballs): The high oddball _encompasses_ the baseline stimulation (i.e. reaches baseline intensity, then goes above and beyond it; baseline+), whereas the low oddball does not (i.e. is not even baseline-). This could mean (among other possibilities) that:*

- *the baseline frequency may be reinforced/more sustained in the high oddball (or vice versa disrupted in the low oddball), which might overestimate low oddball contributions where baseline/oddball contrasts are involved (but see concern II)*

We understand the concern regarding a reinforcement of the baseline frequency in the oddball stimuli. Yet, in our proposed analysis, we only analyze oddball responses and its harmonics which are *not* overlapping with the frequency of the baseline stimulus. Therefore, the frequencies 0.125 Hz (oddball frequency), 0.25 Hz (1st harmonic) and 0.375 Hz (2nd harmonic) will be assessed, which should exclude overlaps with the baseline frequency (0.5Hz).

- *there may be a qualitative difference (the most obvious concern being pain threshold-related, cf Supplementary Figure 2 where VAS dips below pain threshold) between the oddballs that would then make an attribution to simple intensity differences harder.*

We agree with the Reviewer that the change in perception quality might be a confounding factor in our analysis. Yet, given the limitations of the stimulation paradigm, we did not see another way to create an oddball stimulus which is less intense, yet still salient compared to the oddball delivered at a high stimulation intensity. The change in percept is in our opinion the attribute that makes the low oddball salient, i.e. standing out from the other (more intense) baseline stimuli. A description of this rationale has been added to section 2.2.2 of the manuscript.

II. *The second concern is: Why is an oddball paradigm needed at all for this comparison (it is mentioned only in adjunct hypothesis of the Hypothesis Table in Supplementary Materials)? In other words, what function does the baseline frequency have/what does it add in terms of elaborating the differences between high oddball and low oddball (EEG power, or rating)? Removing the baseline frequency altogether, the oddballs would then become mere stimuli, without any loss of internal consistency. There are several possible directions to explain why the oddball might be required (instead of a simple frequency-tagged intensity comparison, which would of course have its own weaknesses but I am not sure how the oddball overcomes it), which should be explicitly mentioned to motivate the oddball paradigm.*

We use the oddball paradigm because we are interested in the effects of the saliency of a stimulus on both the modulation of ongoing oscillations and pain perception. To be able to “stand out from the environment” (i.e., be salient), we believe that the oddball is essential to this goal. If there would be only oddball stimuli, making it a regular series of the same stimulus, there would be no difference between the stimulation peaks itself, and we could only analyze the effects the intensity of the stimulus has (between trials of high and low intensity). Yet, the main goal in this investigation is to

see whether the high oddball will elicit a periodic modulation of ongoing oscillations at its stimulation frequency. The presence of this oddball would illustrate that the brain did pick up on either the periodic intensity or saliency change during the stimulation. Since such an oddball task has not been used before, we cannot assume that this will happen. In a second step, the control condition (low oddball) is used to disentangle whether intensity is the main driving factor behind this modulation, or whether the saliency of the stimulus also plays a role. If we for example will not find a periodic modulation for the low oddball, but we did for the high oddball, it would suggest that the saliency of the stimulus does not play an important role in the modulation of ongoing oscillations. In that sense, the comparison between the oddball responses is secondary to the identifying of the oddball responses itself. This rationale is now clarified in the introduction.

- *Hypothesis Table, Hypothesis 1 and 3a states that the baseline serves _as a positive control_ - but wouldn't just looking at the oddball frequency (or its absence) alone also allow the same conclusion whether frequency-tagging worked?*

Hypothesis 1 and 3a are control conditions, because they test for responses that are well-known and that are the main “identifiers” of the frequency-tagging approach, i.e. that a sustained periodic response will lead to a periodic modulation of the EEG signal. Since we are (to the best of our knowledge) the first group to implement an oddball paradigm in a slow sustained periodic stimulation paradigm using noxious stimuli, we cannot be sure whether we will indeed observe a modulation at the frequency of the oddballs. A periodic modulation at the frequency of the oddball would already show that the oddball paradigm works, as the brain responds differently to these stimuli than to baseline. We therefore refrain from making the oddball-related hypotheses a positive control, since there is the possibility that the oddball does not induce a periodic response.

- *Another section reads "As for the phase-locked response, the difference between baseline and oddball will be calculated for each condition and frequency band. Then, for each frequency band, a paired t-test will be employed to compare the peaks related to .high and .low. If the intensity of the stimulus is the main factor in the modulation of ongoing oscillations, the amplitude for .high will be larger than the amplitude for .low. If saliency is more relevant than intensity, the amplitudes of the oddball in the normal and the control condition will be similar to each other." This seems to suggest that the authors expect baseline to be different between the conditions (if it was identical, it would not need to be considered), but they do not explain why (cf. the following point).*

The analysis of the difference between baseline and oddball allows to control for possible differences that might arise between the baseline and oddball response for each condition. As mentioned previously, this is the first time that a sustained periodic oddball paradigm will be

combined with this specific oddball paradigm, and the possible effect on the baseline frequency are therefore also not clear. Given that the baseline stimuli could be perceived at different levels of painfulness given the condition, it is conceivable that this could have an influence on the modulation at the baseline frequency. Appropriate information has been added in section 2.8.2 (p.17) to motivate the use of the relative amplitudes of the oddballs for the t-tests.

- *Finally, using the oddball paradigm may alleviate some issues I brought up at concern I, but this has not been pointed out clearly/is at best implicit.*

In conclusions, it is unclear to me how the baseline aspect (and vice versa, using an oddball per se) is motivated. There may be good reasons, but they are not stated clearly enough. Does sensory entrainment with an oddball paradigm make the oddball frequency (& harmonics) somehow more robust, or the comparison more sensitive, or anything like that? If so, this crucial aspect should be mentioned as a rationale for using an oddball paradigm to begin with. Or does the oddball modulate the baseline frequencies, and comparing their modulation by high versus low is the actual question here? If so, this aspect may have been omitted.

We appreciate the detailed feedback on our paradigm and hope that our responses to the issues raised above have clarified the matter. The appropriate changes in the manuscript have now been made to motivate the use of this specific oddball paradigm and the control condition more clearly.

= = = = = = = = = = =

Individual minor suggestions

- *I suggest considering time as a predictor to account for the inevitable decrease in attention/vigilance in a not very eventful protocol (during EEG it's simply passive perception over at least 24*80=1920s=32min, albeit with pauses)*

We understand the Reviewer’s concern. However, since the trials are randomized within each block across the conditions (2 stimuli of each condition per block), there should be the same amount of “less vigilant” trials in both conditions, making them comparable. Additionally, the study was designed to include many breaks, which will be used to engage the participant, air out the room etc. to keep subject (and experimenter) vigilant. Finally, the experiment is also divided into 2 separate sessions, to make the duration more tolerable. In our experience, similar kinds of uneventful protocols did not lead to large decreases in vigilance.

- *Rating-peaks (i.e. maxima of continuous ratings) can probably be considered more generously than the proposed 1.35 to 2s, for example from 1 to 3s after stimulus-peak), because the subsequent stimulus-peak's maximum should only be reached at the earliest*

after 2+1.35=3.35s (if I understand correctly); that said, not only Mulders 2020 should be considered but also their own pilot data, where rating-peaks were reached ~1s (not 1.35s) after stimulus-peaks.

We thank the Reviewer for pointing out this inconsistency with our own pilot data. The time-window to detect the rating peaks related to the oddball stimulus was adapted to 1-2.6s after the stimulus peak. This takes into account both the range described in Mulders et al. (2020), as well as the range observed in our pilot for high intensity (1.28 ± 0.17s after peak) and low intensity (2.33 ± 0.19s after peak) oddball stimuli.

- *Is no fixation cross (or some gaze target beyond "keeping it steady") employed?*

In our experience in the lab, having to focus on a fixation cross for a prolonged amount of time (i.e. for 80s) can be quite challenging for some subjects. Allowing some flexibility while instructing them to limit eye movements is a reasonable compromise. Moreover, using the frequency tagging approach, artifacts induced by e.g. eye movements are not influencing the EEG signal as much as in shorter ERP trials since the signal is analyzed in the frequency and not in the time domain. Additionally, with this type of stimuli and analysis, the signal to noise ratio is much higher in comparison to other EEG experiments.

= = = = = = = = = = =

- *Does the research question make sense in light of the theory or applications? Is it clearly defined? Where the proposal includes hypotheses, are the hypotheses capable of answering the research question?*

  Yes. As for "clearly defined", I would call out a few instances that could be clearer, conceptually or linguistically.

  Generally consider checking for language, e.g. confusing sentences like "Painful stimuli emerge from the activity of the nociceptive system which is made to respond to high-intensity and potentially damaging somatosensory stimuli and are therefore inherently salient and to facilitate involuntarily capture of attention (Eccleston & Crombez, 1999)."

  Please use consistent/recurrent/recognizable nomenclature and concepts. For example,
  - sometimes you say "salience or intensity", sometimes "saliency or painfulness", sometimes all three, when you basically refer to the same issue (i.e. distinguishing these facets)
  - sometimes you say "normal" and "control" oddballs or conditions, sometimes "high" and "low"
  - sometimes you write "base", sometimes "baseline" (I think as FoS-Tag)

- *"Based on the assumption that the high oddball will be more salient than the baseline stimuli, we expect that they will be perceived as more intense than the baseline stimuli." is begging the question of the relationship between saliency and intensity (or intensity perception, or pain? unclear)*

We thank the Reviewer for pointing out these linguistic inconsistencies. Throughout the manuscript, we made the distinction between these concepts clearer and rectified wherever the nomenclature was inconsistent. Whereas intensity and painfulness are inherently linked to each other and are often used synonymously in the literature, they do not necessarily overlap (e.g. a stimulus can be perceived as intense, but not necessarily painful).

As defined in the introduction, we regard saliency as “the property of a stimulus that stands out from its environment”. In the case of the high intensity oddball stimulus, it is the “being more intense” that makes it stand out from the less salient baseline stimuli. The perception of saliency and intensity are 2 concepts that are very difficult to distinguish, which is why we added the control condition (using a stimulus with a low stimulation intensity as the oddball). This complex relationship is now discussed in more detail in the introduction section of the manuscript.

= = = = = = = = = = =

Is the protocol sufficiently detailed to enable replication by an expert in the field, and to close off sources of undisclosed procedural or analytic flexibility?

- *Yes, mostly. The experimental procedure could be described more completely, for example, "For each condition (i.e., normal / control), 12 trials will be delivered distributed over 6 blocks of 4 thermonociceptive stimulation trials (Figure 3).": This suggests that breaks between blocks (i.e. quadruplets of counterbalanced High|High|Low|Low oddballs) are self-paced between 2 and 5 minutes, but what about breaks between trials (ITI between the High=>High, for example)? Also please indicate an overall duration in 2.5 not just in 2.1, which should be around 30min for EEG setup, plus 24*80s=1920s=32min for stimulation, plus between 5*2=10mins and 5*5=25mins for inter-block-breaks, plus 6*3*X=Xmin unspecified inter-trial-intervals.*

We improved the details about the duration of the experiment both in section 2.1 as well as in section 2.5.

- *I am having some issues following the calibration procedure. In the manuscript proper, the authors write that subjects should experience pain "throughout the entirety of each trial"/"throughout the 40s trial", elsewhere they write that subjects should "overall" (which I read as "generally but not always") experience pain during the trial (legend Figure 3). The*

latter makes sense considering that the sinusoids always start at 35°C, so _none_ of the trials will ever be painful "throughout their entirety" (even if sensitization sets in immediately after peak 1 and all 35° troughs are experienced as painful - which is also unrealistic -, the first ramp will not be). Do the authors mean to say that "Participants will be asked whether they perceived the stimulation AT THE PEAKS as painful throughout the 40s trial"? Operationalization should be crystal clear here. Relatedly, how are subjects instructed what "painful" means, and is the intention to achieve a certain degree of pain (e.g. VAS8 at peaks), or just painfulness (i.e. at least pain threshold/VAS5 at peaks)?

We completely agree with the Reviewer that the instructions for the staircase procedure should be clarified. As suggested, the aim is to reach a perception of “overall painful”, since the oscillating nature of the used stimulation paradigm does not induce a tonic perception of pain. We will implement a suggestion of the 5 common denominators of pain (pricking, burning) to be clear that all subjects have the same understanding of what “pain” means. Our aim is to achieve a rating that is at least over the pain threshold (VAS 5). This is now further detailed in section 2.3 of the manuscript.

- *I do not understand what is meant by "change in perception of the painfulness twice in a row"? That in two consecutive 40s trials, subjects afterwards report an increase in the 3-category ratings system offered ("no pain overall", "only painful in first half of trial", or "painful throughout the trial")? What if category 2 is never chosen, or the second half of the trial is painful?*

We apologize for the unclear formulation of the instructions. We indeed meant to say that in two consecutive trials, the subject changes the category of the answer. Usually, in the beginning, the trial is not painful. We then increase the temperature of stimulation step by step until the subject says that they perceived the peaks of the full 40s as painful. We then lower the temperature by 1°C. If the subject then changes the category to “only painful in first half” we consider the higher temperature as the temperature of stimulation. Conversely, if the subject says that the trial still feels painful for the full duration, we will decrease the temperature of stimulation until the category of perception changes.

Due to the habituation that inevitably occurs during the 40s of stimulation (and which could be clearly seen in earlier pilots for this experiment), no subject reported that the pain perception increased in the second half of the stimulation in the pilot. We therefore decided to not include this as a category for the staircase procedure. The second category (“painful only in first half”) has been shown in our pilot studies to be the most chosen category of perception. Due to the very particular stimulation paradigm and duration, the stimuli are usually perceived as intense at the beginning and then lose some intensity throughout the trial.

= = = = = = = = = = =

Is there an exact mapping between the theory, hypotheses, sampling plan (e.g. power analysis, where applicable), preregistered statistical tests, and possible interpretations given different outcomes?

Yes.

= = = = = = = = = = =

For proposals that test hypotheses, have the authors explained precisely which outcomes will confirm or disconfirm their predictions?

Yes, mostly. I was wondering if 4b and 6b are meant to be positive controls, as well - if frequency-tagging works with the high oddball but not the low, surely there is nothing to compare?

The use of the “low oddball” condition serves as a control for the “high oddball” condition. The response at the high oddball cannot disentangle between intensity and saliency, since the oddball stimulus incorporates both properties. The control condition on the other hand should be perceived as salient (because it is “standing out” from the baseline stimuli), but not as more intense than the baseline stimuli. Thus, if the frequency-tagging does not work with the low but does work with the high oddball, the results would indicate that saliency does not contribute as much to the modulation of ongoing oscillations as intensity does. The comparison of the amplitudes of the oddball responses is only a secondary step to elucidate the relationship between intensity and saliency more closely. The explanation in section 2.8.2 (p.17) has been adapted to reflect this argument more clearly.

= = = = = = = = = = =

Is the sample size sufficient to provide informative results?

Yes.

= = = = = = = = = = =

Where the proposal involves statistical hypothesis testing, does the sampling plan for each hypothesis propose a realistic and well justified estimate of the effect size?

Yes.

= = = = = = = = = = =

Have the authors avoided the common pitfall of relying on conventional null hypothesis significance testing to conclude evidence of absence from null results? Where the authors intend to interpret a negative result as evidence that an effect is absent, have authors proposed an inferential method that is capable of drawing such a conclusion, such as Bayesian hypothesis testing or frequentist equivalence testing?

Yes, well, mostly. I am having some issues with the way some interpretations are phrased concerning whether hypotheses are rejected or not. For example, "A similar amplitude of the oddball in the high and low oddball condition would show that the oddball response is mainly driven by the saliency of the stimulus. If the oddball in the low oddball condition would lead to a smaller response compared to the oddball in the high oddball condition it would indicate that the oddball the intensity of the stimulus is responsible for the peak related to the oddball." is quite assertive as to the mechanisms involved. I would prefer more epistemologically cautious phrases like "supports a role of" or "suggests that XYZ is not reflected in the signal" or so. Saliency and intensity are not the only thinkable contributors to either readout of this study.

We thank the Reviewer for bringing up this concern and agree that the current phrasing might be too deterministic. We adapted the phrasing accordingly in the different sections of the hypotheses table.

= = = = = = = = = = =

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.

Yes.

= = = = = = = = = = =

Have the authors clearly distinguished work that has already been done (e.g. preliminary studies and data analyses) from work yet to be done?

Yes. (as a minor aside, I was wondering, why do the ratings in Suppl Fig 1 go up again at the end of the trial?)

The ratings increase at the end of the shown time window in Suppl. Figure 1 because the rating was given *after* the 80s of stimulation shown in the figure on the x-axis- Since there is a 1-3 second delay between the peak of the stimulation and the peak in the rating of perception, the time window of analysis will have to be extended to ~83s to include the ratings of the 10th oddball.

= = = = = = = = = = =

Have the authors prespecified positive controls, manipulation checks or other data quality checks? If not, have they justified why such tests are either infeasible or unnecessary? Is the design sufficiently well controlled in all other respects?

Yes they have specified positive controls.

= = = = = = = = = = =

When proposing positive controls or other data quality checks that rely on inferential testing, have the authors included a statistical sampling plan that is sufficient in terms of statistical power or evidential strength?

Yes.

= = = = = = = = = = =

Does the proposed research fall within established ethical norms for its field? Regardless of whether the study has received ethical approval, have the authors adequately considered any ethical risks of the research?

That seems to be for the ethics committee to decide, who have given a positive vote. However, for me, some concerns remain regarding stimulus intensities. These are possibly device-related (non-)issues, but important enough to emphasize: The temperatures mentioned are very high. If tissue actually reached such temperatures, burns would be inevitable in the timeframes employed, especially considering the flat addition of 3°C (safe exposure duration decays exponentially by intensity, cf. Moritz & Henriques, Am J Pathol 1947); actual temperatures of 50°C would be excruciating for 90+% of people. The authors have piloted the procedure and it seems to work, but I would suggest they figure out the discrepancy to established temperature ranges.

We agree with the Reviewer that the tolerability of the painful stimuli is an important aspect to consider when planning an experiment involving thermonociceptive stimuli. As mentioned, we did pilot this specific paradigm and it was tolerated quite well by participants. It should be mentioned that the particularity of this paradigm is the oscillating pattern of the stimulation, which makes thitype of stimulus much less painful compared to a tonic stimulation at the peak stimulation temperature. Here, the peak temperature is only reached for a very brief time, before the stimulation oscillates back to the non-painful) baseline temperature of 35°C. Moreover, to avoid habituation or sensitization, the thermode will be displaced after each stimulation Finally, as the stimulation temperature will be adjusted to the individual, we are confident that we will be able to deliver painful, yet tolerable stimuli in this experiment.

Review by Markus Ploner, 09 Oct 2023 15:48

The proposed study aims to investigate the effects of stimulus intensity and saliency on modulations of pain-related neural oscillations. To this end, the authors propose a paradigm in which sustained periodic heat stimuli are applied to 30 healthy human participants. In an oddball paradigm, deviant stimuli of higher or lower intensity will be interspersed. Differences in amplitudes of oscillations between high and low oddball stimuli would indicate that oscillations are influenced by stimulus intensity. In contrast, a lack of a difference would be taken as evidence for an effect of saliency on oscillations.

The study is well-planned, and the manuscript is mostly clear and convincing. However, the framework, the paradigm, and the analysis might benefit from modifications and added details.

We thank the Reviewer for his concise and constructive feedback. We have now taken the Reviewer's comments into consideration and believe this will further improve our work.

1. Framework. The framework and the analysis propose a binary view of stimulus intensity and saliency effects on pain-related brain activity. Brain activity is either influenced by stimulus intensity or saliency. However, considering previous evidence for saliency and stimulus intensity effects on pain-related brain activity, a mixture of both effects appears possible, if not likely. So far, a difference between high and low oddball stimuli will be interpreted as a stimulus intensity effect but does not provide any information about possible additional saliency effects. In contrast, a lack of a difference between high and low oddball stimuli will be interpreted as an exclusive saliency effect. Thus, in the likely case of a difference between high and low oddball stimuli, the study's outcome would be evidence for a stimulus intensity effect but essentially no information about saliency effects. The authors might fundamentally re-consider their framework and analysis so that it can account for and quantify non-binary effects of saliency and stimulus intensity.

We agree with the Reviewer that the effects that we are likely to observe will not be completely attributable to either saliency or intensity. It was also not our intention to consider them a dichotomy. Upon re-assessing the manuscript, we realized that our wording of the hypotheses was not ideal, and we adapted both the introduction as well as the sections 2.8.2 and 2.8.3. with a summary of the justification below.

We are indeed not trying to completely isolate either saliency or intensity from each other but rather aim to identify how they (and their potential interaction) influences the modulation of ongoing oscillations. We would like to emphasize that we want to characterize the influence that saliency has on the modulation of ongoing oscillations. To this end, we compare whether a highly intense and salient oddball and a low intensity (but still salient) oddball stimulus are able to elicit a periodic modulation in the EEG spectrum. Whether we are able to elicit a response using these different

conditions is the core of this investigation. Given previously published investigations using oddball paradigms (Rossion et al., 2020), we can expect to see a modulation at the frequency of stimulation for the highly intense and salient stimulus. Yet, it is not that clear whether this will also be the case for the low intensity oddball stimulus. Importantly, observing a modulation at the oddball frequency in this paradigm would already inform us about the saliency of the stimulus, as this would most likely be the main factor driving that modulation (given the low intensity of the stimulation) and without which there would presumably be no significant change at the frequency of the low oddball stimulation.

Only in a secondary step, after confirming that both oddballs elicit a periodic response, will we compare the amplitudes of these oddball responses. While we will not be able to clearly quantify the influence of either saliency or intensity on a given amplitude, this analysis will nevertheless inform us about potential ratios in which these two factors interact for a specific oddball stimulus.

Finally, we would like to mention that saliency and intensity are concepts which are inherently linked, as a highly intense stimulus is frequently also highly salient, and therefore rather difficult to disentangle. Additionally, in most experimental settings we have encountered so far, they appear intertwined. Thus, the information gathered in this experiment will be particularly useful in the comparison of stimuli of different modalities, which might vary either in saliency or intensity as compared to the thermonociceptive stimulus. Our results will hopefully make it easier to understand which properties of this stimulus had a larger impact on the observed modulations and will guide us in creating experimental setup with different stimulus modalities that are equal in their potential to elicit a periodic brain response.

2. Paradigm. The paradigm includes a sustained periodic heat stimulation with high and low oddball stimuli. However, high and low oddball stimuli do not only differ concerning stimulus intensity but also concerning painfulness. The high oddball stimulus is always painful, whereas the low oddball stimulus is always non-painful (Supplementary Figures 1 and 2). Thus, the quantitative difference in saliency between both conditions is confounded by a qualitative difference between painful and non-painful stimuli. This should be carefully considered and accounted for. The authors might consider adjusting the paradigm so that all stimuli are painful.

The difference in qualitative perception might indeed be a confound in this experimental setup. Yet, there are multiple reasons for our choice of stimulation temperatures, both conceptual as well as practical:

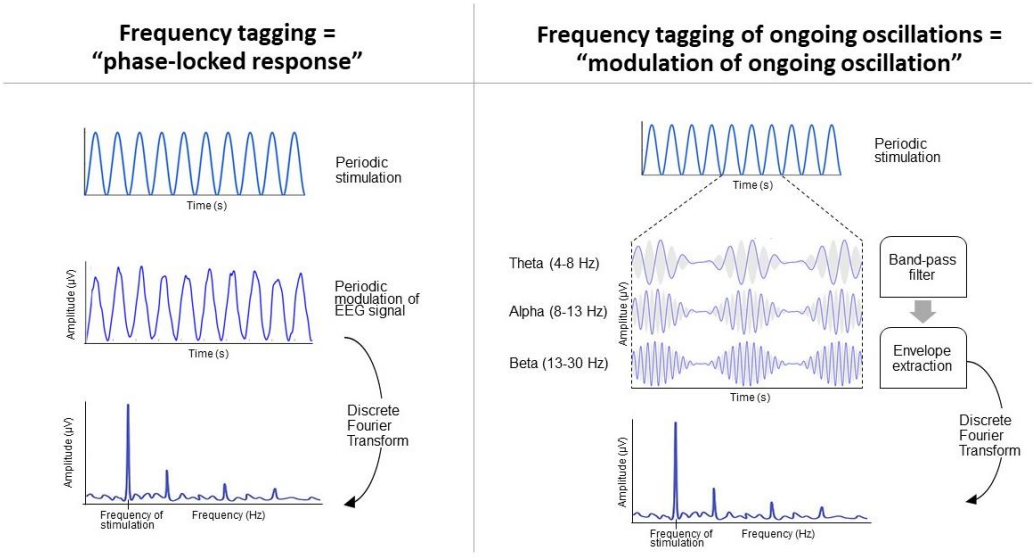
Our main aim was to create a paradigm, for which the baseline stimulus would be perceived as painful, and to assess how the salient deviations from this stimulus (both to a higher and lower stimulation intensity) would modulate the EEG signal, and whether these modulations correspond

to the changes in perception of the stimulus. For this purpose, the change in qualitative perception of the stimulus seems to be a natural consequence of the change in intensity. Furthermore, to make a less intense stimulus salient, it needs to be quite different from the preceding stimulus to “capture attention”. Using a qualitative different stimulus allows us to create this percept. A summary of this rationale has been added to section 2.2.2 of the manuscript.

More practical concerns relate to our aim for stimulation temperatures that would be clearly distinguishable from each other to allow participants to adequately trace the stimulation pattern during the VAS rating session. In our pilot experiments for this project, we saw that temperatures that were too close to each other were not distinguishable for most participants, and they perceived a tonic stimulation instead of the sustained periodic pattern. Moreover, the feasibility had to be considered for the choice of temperatures. With the chosen temperatures, the high intensity oddball is already perceived as very painful. If we had to increase the stimulation temperature of the lower oddball, the temperatures for baseline and high intensity oddball would consequently also raise. Given our experiences during the pilot phase, we believe that that would lead to conditions in the experiment that would not be tolerable for most participants.

3. Writing. The general reader might need to become more familiar with the frequency tagging approach and the underlying logic. A more straightforward explanation of the approach with an intuitive figure showing what will be analyzed and termed “phase-locked response” and what will be called “modulation of ongoing oscillations” would be helpful.

We understand that the frequency-tagging approach might not be intuitive for a naïve reader and have therefore adapted a figure to illustrate both the general frequency tagging (i.e. analysis of the phase-locked response) as well as the frequency tagging of ongoing oscillations (i.e. analysis of the modulations of ongoing oscillations). The figure has been added in section 2.7.1. We hope that this figure will aid the reader in understanding the different analyses.



4. Analysis of behavioral effects. The analysis window for pain ratings is 2 sec after the stimulation peaks. However, the pilot experiment has revealed that pain ratings in the low oddball condition peaked after 2.33 sec. Thus, an adjustment of the analysis window might be appropriate.

We thank the Reviewer for pointing out the discrepancy between our suggested analysis window and our own pilot data. We adapted the time window to identify the peak of the rating between x to 2.4 seconds after the onset of the oddball stimulus. These values reflect the data we captured in the pilot experiment and still incorporate the response time window reported in (Mulders et al., 2020).

5. Analysis. The definition of the frequency bands does not adhere to the COBIDAS recommendations (Pernet et al., Nat Neurosci, 2020). This should be corrected.

We were not aware of this general recommendation but are happy to adhere to this suggestion to homogenize the definition and analysis of ongoing oscillations. We changed the definition of the frequency bands to the recommended: theta (4-8 Hz), alpha (8-13 Hz) and beta: 13-30 Hz. Accordingly, we changed the 4^{th} order Butterworth filter used at the beginning of the pre-processing pipeline to filter from 0.05 – 30 Hz instead of 0.05 - 40 Hz.

6. P.15, second paragraph. In the second last line, "high oddball condition" might have been confounded with "low oddball condition."

We thank the Reviewer for pointing out this mistake and have corrected it accordingly.

References

Mulders, D., de Bodt, C., Lejeune, N., Courtin, A., Liberati, G., Verleysen, M., & Mouraux, A. (2020). Dynamics of the perception and EEG signals triggered by tonic warm and cool stimulation. *PLOS ONE*, *15*(4), e0231698. https://doi.org/10.1371/journal.pone.0231698

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard Power as a Protection Against Imprecise Power Estimates. *Perspectives on Psychological Science*, *9*(3), 319-332. https://doi.org/10.1177/1745691614528519

Rossion, B., Retter, T. L., & Liu-Shuang, J. (2020). Understanding human individuation of unfamiliar faces with oddball fast periodic visual stimulation and electroencephalography. *European Journal of Neuroscience*, *52*(10), 4283-4344. https://doi.org/https://doi.org/10.1111/ejn.14865