I appreciate the detailed reply. The conclusions do have to be justified appropriately by the results however; and at the moment they still do not line up.

We thank the Reviewer for the follow-up on the previous response. We have adapted the manuscript accordingly (mainly the Design Table). Since the Sample Size section has grown to be quite lengthy, we chose to move it to the Supplementary Materials (at the end of the manuscript in Stage I submission) to enhance the readability of the article.

1) **The authors should indicate what effect size their N=40 does give them power to detect; then judge if a smaller effect would be theoretically interesting. If so, they should be clear in the Design Table that a non-significant result does not indicate the theory has been shown wrong.**

Estimating effect sizes for Linear Mixed Models (LMMs) is not a straightforward process, and there is still much debate about the ideal procedure. Since the power calculation is based on a simulated data set, our first approach was to repeat the simulation, but with 40 instead of the initial 12 participants of the original data set. The means and standard deviations stayed the same. Based on this simulation, a LMM was calculated, and the F-values transformed to partial η^2 values. This resulted in an estimated effect size of 0.09 for the interaction between cue and temperature for the phase-locked response in the EEG signal. We thus estimate to detect a slightly larger effect size than found in the study of Mulders et al. (2020) ($\eta^2_p = 0.06$ for the interaction in the phase-locked response). Therefore, we would theoretically still be interested in a smaller effect than found in the simulated model, which means that a non-significant result will not necessarily indicate that our hypotheses was wrong. This rationale is now indicated in the Design Table as well as in the Sample Size rational section (pp. 22-24). Moreover, we will consider conducting post-hoc power calculations in case of non-significant results, with the aim to assess the actual power that we reached with the acquired data set.

Additionally, the software "G*Power" (V. 3.1.9.7) (Faul et al., 2007) was used to calculate the required effect size to reach specified power (0.98), alpha error probability (0.02) and the sample size values (n=40). Unfortunately, the software does not offer this calculation for LMMs, and we were not able to find any adequate alternative. Instead, we approximated the model using the calculation for a repeated measures ANOVA, with within factors only. 1 group was compared along 4 measurements, with a correlation among repeated measures of 0.5 and a nonsphericity correction of 1. This resulted in a calculated required effect size f= 0.282, which equals an η^2_p of 0.074. This calculation shows that the targeted sample size should allow us to detect at least effect sizes similar to previous investigations. Since the ANOVA does not take the variance within subjects into account and is therefore less robust than an LMM, we can expect to be able to detect even smaller effect sizes with our model.

2) **The authors still need to indicate what power they do have for each test in the Design Table. They could proceed in the same way; for their planned sample size, indicate what effect they do have sufficient power to detect (for THAT test), and if smaller would still be interesting, appropriately indicate in the final column of the table that a non-significant result would not refute the claim tested - including for outcome neutral tests, this means toning down the existing language.**

The software G*Power was used to calculate the effect size required for the Wilcoxon signed rank test, which will be used to identify amplitudes at the frequency of interest which are significantly larger than zero. Specifically, the sensitivity of a one-tailed one sample case Wilcoxon signed rank test with a normal distribution and an alpha error probability of 0.0125 (corrected for multiple testing), a power of 0.9 and a sample size of 40 participants was assessed. This calculation estimated an effect size of d=0.6 which equals an $\eta^2_p$of 0.083 (medium effect). Since the amplitudes of the modulation of ongoing oscillations tend to be rather small, especially for the theta frequency band, we will also be interested in effects that are smaller than η^2 =0.083. As for the previous analysis, we will consider a post-hoc power calculation to verify that the targeted power has been reached and to assess whether potential non-significant results might be due to an insufficient sample size.

Large effects for differences in perception of painful stimuli using expectation cues were found in previous investigations (effect sizes approximated from reported test- statistics) (Atlas et al., 2010 ; Hauck et al., 2007). This is supported by our pilot data, which showed very clear differences (similar percentage change between conditions as in Atlas et al. (2010)) between conditions HM and LM. We therefore expect to find at least an intermediate effect size in our data. Since structurally, the same LMM will be used for the estimation of the effect of cue and temperature on pain ratings as for the effect on the amplitude of the EEG signal, we can assume that our sample size will be sufficient to inform about effects of temperature and condition on pain ratings. Therefore, a smaller effect would not be interesting. G*Power was used to calculate the sensitivity (required effect size) for the power, alpha error and sample size associated with the LMMs of the ratings, using a repeated measures ANOVA, within factors, as conservative approximation for the LMM. The estimated effect size corresponds to η^2=0.058.

3) **Bayes factors confront a similar issue: They are only meaningful tests of a theory, if the scale factor represents the sort of effect predicted by the theory. Note in this case, the most relevant aspect of the prediction of a theory is not the minimal meaningful effect, but the sort of effect predicted. The rough size of effect predicted is in general easier to justify scientifically than the minimal meaningful effect; but still there needs to be a justification. A default is just a suggestion to consider if e.g. a Cohen's d of 0.7 is actually relevant; it is not to be used without thought. Rather than mix inferential systems, however (in this case frequentist and Bayesian), the easiest thing to do here would be to rely on the frequentist stats for inference, so these are what appear in the Design Table. There would be no harm in reporting default BFs for information for the reader - then no justification of the prior (model of H1) is needed - but the authors stick to the rationale of hypothesis testing with power. In that case, if the study is not powered to detect small but interesting effects then this is simply recognized i nthe conclusions afforded by the analysis.**

As detailed in the previous responses, we added disclaimers in the Design Table that we will not be able to detect the smallest still interesting effect with the recruited sample size.

We understand that the addition of the Bayesian post-hoc analysis only has limited usefulness within the analysis of this study (similar to the limitations of the frequentist approach). Yet, we would like to keep the post-hoc Bayesian hypothesis testing in the analysis section, as this was added in response to a Reviewer who is now not involved in the process anymore. We believe that it would be suboptimal to make changes to the previously agreed upon form of the manuscript at this point.

4) **The alternative is go over to BFs for all tests as the system of inference, but this would involve thinking everything through again, which is why I say the simplest thing is to go with the power and N such as they are (and as the authors say they are good by field standards), but recognize the inferential consequences of this in the Design Table.**

In sum, what is needed is every conclusion is scientifically justified by the inferential procedure in every row of the Design Table in an explicit way.

We agree with the Reviewer that at this stage it does not make sense to introduce a complete change of the analysis (i.e., swap completely from frequentist to Bayesian analysis approach). We hope that the changes made to the Design Table as well as the additions to the Sample Size section (p.xx) are sufficient to convey the fact that despite our best efforts, we are testing rather for the expected than for the smallest still interesting effect.

References

Atlas, L. Y., Bolger, N., Lindquist, M. A., & Wager, T. D. (2010). Brain Mediators of Predictive Cue Effects on Perceived Pain. *The Journal of Neuroscience*, *30*(39), 12964-12977. https://doi.org/10.1523/jneurosci.0057-10.2010

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

Hauck, M., Lorenz, J., Zimmermann, R., Debener, S., Scharein, E., & Engel, A. K. (2007). Duration of the cue-to-pain delay increases pain intensity: a combined EEG and MEG study. *Experimental Brain Research*, *180*(2), 205-215. https://doi.org/10.1007/s00221-007-0863-x

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