San Sebástian, 17/11/2023

Dear Chris Chambers,

Thanks for the revision and the constructive comments from you and the reviewers. We have implemented many of the suggestions proposed by the Reviewers or justified our decisions and rationale in this response letter.

It follows a point-by-point response to the Reviewers' concerns.

We hope the manuscript is deemed suitable to receive the Stage 1 IPA and to hear from you at your convenience.

Best Regards,
The #EEGManyLabsMathewson replication Team

Review #1 PCI-RR

Reviewed by Chris Allen

This is one of the best pieces of work I have been asked to review!

It aims to replicate an important study with foundational implications for influential theories of oscillatory brain function. There have been previous replication attempts in this area, including a Registered Report, but I am not aware of such a multi-lab attempt with potentially such high power (but see point 4 below). The outcome of this, I expect, will be of great interest to the related field.

Below I list some suggestions for adjustments in order of appearance in the manuscript, none of which I see as fundamentally critical.

We thank the Reviewer for their encouraging comments on our Stage 1 RR.

1. Title: would it be possible for the title to be a bit more specific and informative? Something along the lines of 'Testing the replicability of alpha phase determining visual perception'. I realise that's not as catchy, but I found the current title a little vague.

We understood the Reviewer's suggestions and decided to change the title. Please note, however, that our RR describes a replication study (Study 1) and a new experiment (Study 2), which we will run only if the results from Study 1 are positive. We think that this aspect of the RR should be reflected in the title. Furthermore, we would like to clarify the final title of the manuscript (Stage 2) might change to include the results of Study 2, should this be run.

The new title is **An #EEGManyLabs study to test the role of the alpha phase on visual perception (a replication and new evidence)**.

2. The third sentence assumes subjective experience is a continuous flow, but phenomenology suggests this is not straightforwardly the case (see e.g., Husserl, *The Phenomenology of Internal Time-Consciousness, or Busch and VanRullen, 2014, Is visual perception like a continuous flow or a series of snapshots? In Subjective Time, MIT Press, or Dainton, 2010/232023, Stanford Encyclopaedia of Philosophy, Temporal Consciousness). This could be simply resolved by losing the first part of that sentence.*

The Reviewer touched on an interesting and potentially long-lasting point of discussion. To resolve the issue, we decided to accept the Reviewer's suggestion and delete the first part of the sentence.

3. In the hypothesis at the end of the introduction, I thought the statement under a negative finding that "there is no evidence for visual perception to operate in cycles" was a little strong, and perhaps should be rephrased to refer to this study's evidence. This also relates to the point 9 below. Relatedly, I thought it would be potentially more informative if the decision criteria at the end of the introduction (which seems primary) were based on the outcome of the Bayesian analysis rather than the frequentist statistics.

We agree with the Reviewer that any evidence challenging the significance of the alpha phase must be linked to the study design. For this reason, the full sentence reads as follows: "If behavioural and/or electrophysiological measures have no relation to the phase of the pre-stimulus, low-frequency brain rhythms, we should conclude that we do not replicate Mathewson et al., (2009) results; therefore, within the studied parameters, there is no evidence for visual perception to operate in cycles."

For the decision criteria, we are following the criterion that has been agreed upon in the #EEGManyLabs position paper (#EEGManyLabs: Investigating the replicability of influential EEG experiments; https://doi.org/10.1016/j.cortex.2021.03.013) and therefore, success is defined as a statistically significant random-effects meta-analytic estimate (at p < .02) combining the results from the different laboratories. We will stick to this decision rule for deciding whether study 1 was a replication success or not. However, in line with the reviewer's comment, we also plan to run corresponding Bayesian analyses in JASP for all statistical tests. Bayes factors (BF) larger than 6 will be considered as evidence for an effect and BF smaller than 1/6 will be considered as evidence against an effect. BF between 1/6 and 6 will be considered as undecisive. The random-effects meta-analytic estimate will be complemented with a random-effects Bayesian meta-analysis. According to the result of this analysis, we will run study 2 if study 1 provides inconclusive evidence or evidence for an effect. This was unclear and we have added a sentence in the "Meta-Analysis" section and also in Table 2 legend.

4. I found the first and second sentences of the participants section a little ambiguous. It could be read that the 7 labs will collectively provide a 35 participant data set, i.e., 5 each, OR, each lab would aim to contribute a 35 participant data set. My confusion was not helped by other sections seeming to imply that both the full data set had n=35 (e.g., "final sample size (N=35)"), and each lab is to produce a complete data set (e.g., "compute effect sizes (Cohen's d) for each individual lab"). If it is the former, and given the reasons well specified in the introduction for effect sizes reported by Mathewson et al., being larger than we might expect in replication, *then I do not think the study as it stands should be described as "high-powered". If there is a real effect but it is smaller, as the authors suggest is likely, the study would be underpowered. Would it be possible to recalculate power estimates based on smaller effect sizes? Perhaps, based on the estimates of VanRullen et al., 2016 described in the introduction? I appreciate the intention behind the use of the Mathewson et al., effect sizes, but if they are used, I suggest explicitly describing the limitations in interpreting a negative outcome (also see point 9).*

If, however, each lab is to contribute a 35-participant data set, then great, and I think this should be made clearer and the total minimum n should be stated. It might also help if the smallest effect size reliably detectable was stated. Inclusion of such should be informative either way. I also wondered whether a simpler concatenation of the data across labs might complement the meta-analytic approach but be slightly more powerful. I also thought it might be possible to combine evidence across labs more efficiently by taking advantage of Bayes Factors being transitive (Morey and Rouder, Psychological Methods, 2011).

Thanks for spotting a possible source of misunderstanding in our RR. We have made it clear that each Lab (9 in total) will contribute to the replication with the recordings from 35 participants each for Study 1. For Study 2, 4 Labs have committed to record 35 participants each. Please also note that conclusions from Bayesian statistics will complement frequentist statistics and will be useful in particular for cases of non-significant results, because the BF analysis will tell us whether our data was inconclusive or whether we even found evidence against an effect. The meta-analytic BF will also decide whether we will run Study 2. We won't be concatenating data because we want to quantify the variation in effect sizes across labs, to understand if the variability exceeds the amount expected as a result of measurement error (Pavlov et al, 2021).

5. I didn't see the need to restrict the age of participants to a 12-year range.

Because Study 1 is an exact replication of another study, we decided to use the same participants' age range as in Mathewson et al., 2009. Furthermore, we know that the alpha frequency and power change with age, so we believe that opening the age range too wide would introduce noise in the data and make the interpretation of the results complex, especially if we fail to replicate the original study. Finally, 18-30 years old is an age range widely used in the literature, therefore making our dataset comparable with others.

6. I thought there should be a sentence qualifying the timings as approximate, or specifying two sets of timings, as the refresh rates of the monitors used across the different labs would mean these exact timings cannot be followed by all labs.

Thanks for this comment. We realised that we have not specified that each Lab must check the precise timing of the experiments by using an oscilloscope or a photodiode before starting the experiment as a requisite in the #EEGManyLab project. We have introduced this information in the manuscript. Furthermore, timings have been converted into numbers of frames with a 100Hz monitor, which is the most used in our replication (see Table 3). Finally, figure 1 has been updated to match the timings using a 100Hz monitor.

7. The sentence "The experimental session counts 16 blocks of 72 trials each", should that be "The experimental sessions consist of 16 blocks of 72 trials each"?

*Thanks. The sentence has been changed to "*Within each study, the main experimental session consists of 16 blocks of 72 trials each."

8. Should an impedance target be prescribed? I understand this can be equipment-dependent, but then maybe it could be lab-specific.

The quality check of each dataset will be estimated based on the signal-to-noise ratio (SNR) at the N1 component, as stated in the Quality Checks section. An impedance target would be problematic because different equipment will be used, and impedance comparison across systems can be hard. Furthermore, it will open to other difficult decisions, such as how many electrodes/trials will be allowed to cross the target before discarding the dataset. We therefore considered that the SNR will constitute a better measure of the dataset quality.

9. I would have liked to see more weight given to adjudication between support for null vs. lack of evidence in case of a negative finding in the main text. I think differentiating between these two potential outcomes could be important information for the field. I understand that the Bayesian meta-analytic approach offers the potential to do this, so I was wondering whether the Bayesian analyses could receive a greater emphasis in terms of decision criteria. For example, at the end of the introduction. This might mean using Bayesian equivalents for the primary t-tests (which can be derived from T statistics) and recalculating the sample size estimation based on Bayesian simulations, but in my experience, the outcome invariably aligns between Bayesian and frequentist estimations. Alternatively, frequentist statistics can assess equivalence (Lakens et al,. 2018, AMPPS). This relates to point 4 where to fairly assess equivalence, a smaller expected effect size may be required.

We thank the Reviewer for this important aspect of the RR. We would like to emphasize that the final conclusion about whether or not the null hypotheses will be rejected, thus whether the replication was successful, will depend on the frequentist statistics because this decision rule has been set in the already approved and published position paper of the #EEGManyLabs consortium. A deviation from this rule for our study would introduce inconsistencies to the global replication project. (See also our response to point 4.) The frequentist statistics will be complemented by corresponding BF analyses, which will be particularly informative in cases where the null hypothesis cannot be rejected. As stated in the text, and in Table 2, Bayesian statistics will be used to inform any non-significant results and to establish whether we will run Study 2 (please see our response to point 4). Null-hypothesis testing and frequentist statistics are used because our first aim is to replicate Matthewson et al., 2009, including their original analysis, which has also informed the estimation of the sample size. In the hypotheses section, we have stressed this point a bit further. Please note that we would like to clarify that with the benefit of current knowledge, many design choices for testing the role of the alpha phase in perception would be different today. However, this is a Stage 1 RR whose main goal is replicating a very influential study in the field. This argument will be part of our discussion in the Stage 2 phase.

10. Related to the above point and based on the expectation of small effects described in the introduction, I thought the prior applied in the Bayesian analyses should probably reflect this aspect of the hypothesis and be smaller than the default 0.707 scaling factor (see Dienes, 2011, Perspective on Psychological Science).

Although we agree with the reviewer that the effect size might be smaller than 2, it is difficult to commit to a more informed, narrower prior distribution without having strong evidence on the size of the effect we are looking at. However, as discussed by Berkhout et al., 2023 (doi: https://doi.org/10.3758/s13428-023-02093-6), Bayesian meta-analysis seems to be robust against the choice of the distribution parameters. Wider/narrower prior distributions will result in a larger/smaller Bayes factor, respectively, but these differences are small and do not change the substantive conclusions.

11. Table 1 describes decision criteria based on $p<0.05$, whereas the text describes $p<0.02$ criteria (also true in the final table). Can this be checked or reasons for the discrepancy given? Also, I did not find the description of Hp b1 in this table very easy to understand.

There was indeed a mistake. Thanks for spotting it.

12. Table 2 and the final table seem to repeat much of the same information, I realise the final one conforms to guidelines, but I still wondered whether they could be integrated. I also thought the final table would benefit from basing criteria on Bayesian tests to more straightforwardly differentiate between negative and inconclusive outcomes, and this might help align the tables.

The Reviewer is right; the final table repeats information already present in Tables 1 and 2. Please note, however, that Tables 1 and 2 will be part of our final manuscript. The final Table is a requirement from the journal, which will be deleted.

Review #2 PCI-RR

Reviewed by Luca Ronconi

We thank this Reviewer for the constructive suggestions.

1) I think that sometimes the description of the rationale Study 2 is overly simplistic. It is true that in Study 1 subjects will have strong temporal expectation becuase of the very short and fixed inter-stimulus interval (ISI) employed, but such expectation will be very likely to be present also in Study 2, altough in an attenuated form. This should be more clearly stated.

We agree with the reviewer that claiming no temporal expectation at all in the second experiment may be an overstatement. We have modified the text accordingly. We explain that we have selected ISIs drawn from a distribution that minimizes temporal expectations.

Please note, however, that we have selected the exponential distribution to generate the ISI (Figure 1A) because this is the distribution that has a maximum entropy and it is the only one with a constant hazard rate (see http://dx.doi.org/10.11568/kjm.2013.21.4.429 and https://doi.org/10.1007/978-1-4419-6646-9). Therefore, although subjects may have temporal expectations (no stimulus will be presented at times below 400 ms or times above 5000 ms), the temporal expectation will be distributed evenly in the 400 ms to 1000 ms interval. In other words, the hazard rate in the interval 400 ms to 1000 ms will be constant (Figure 1B), contrary to the hazard rate obtained for a uniform distribution, which increases in the interval to a peak value of 1 at 1000 ms (Figure 1B) and decreases sharply at longer time intervals.

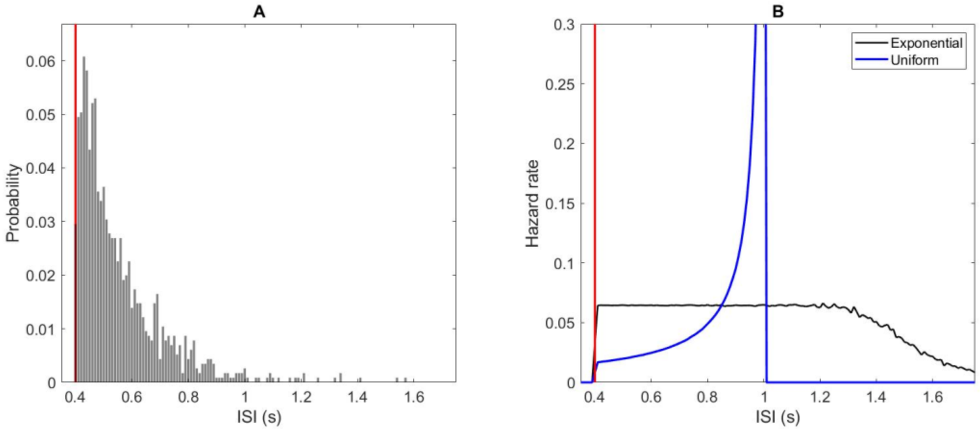

Figure 1: (A) ISI probability distribution for the distribution used in Study 2: 400 ms plus a variable interval extracted from an exponential distribution with mean 150 ms. (B) Hazard rate for the distribution used in Study 2 (black), compared to the hazard rate of times distributed uniformly between 400 and 1000 ms. Red vertical line indicates the ISI in Study 1. In both graphics, we plot the average of 10000 simulations of 1152 trials each (the number of trials of the replication).

2) I invite the Authors to consider that, in case Study 1 will be succesfully replicated, temporal expectation is only one out of the different aspects that differentiate the two studies. For example, the evoked response and the concomitant 'phase reset' created by the appearence/disappearence of the fixation cross is likely to be more influential in Study 1 than in Study 2, becuase of the increased temporal distance between events present in Study 2. How will the Authors take into account this factor in Study 2 to be sure about their conclusion on the role of temporal expectation? This should be better clarified.

The use of the exponential distribution to generate ISIs has several advantages. On one side, it minimizes the chances of subjects to allocate their attention to specific time points, on the other, it is possible to select the parameters to maximize the number of ISIs that falls within a specific time window. In order to increase the number of trials appearing at latencies comparable to Study 1, we have modified the parameters of the exponential distribution, selecting a mean value λ equal to 150 ms (see Figure 1a in this response letter for a distribution of the ISI obtained with this set of parameters). With this choice, 3.5% of the trials would be presented at 400 ms (analogous to Study 1) and 25% of the trials would be presented at ISIs shorter than 450 ms (half alpha cycle).

Review #3 PCI-RR

Reviewed by Alexander Jones

We thank this Reviewer for the constructive suggestions.

The prediction is that if temporal expectation may be present in study 1 with the fixed ISI but not in Study 2 which includes a variable ISI. I think one thing to consider is any potential influence of the foreperiod effect/hazard function which may be present in Study 2. That is, there is some type of temporal expectation (the effect of time) in Study 2 which is not there in Study 1. So the question is whether you have considered if the foreperiod/hazard effect can influence the results? Would be good if you can add something regarding this.

The Reviewer is right that the manipulation used in Study 2 will have an effect on hazard rate, which is an important point to mention in the manuscript. However, we choose to generate ISI in Study 2 using an exponential distribution, which is the only distribution with a constant hazard rate (http://dx.doi.org/10.11568/kjm.2013.21.4.429 and https://doi.org/10.1007/978-1-4419-6646-9)

Why pick Pz and Fz as the analysis electrodes? I get that these were the ones used in the original paper, but in that paper they also used Oz and show an effect on P1/P2 components.

We agree with the Reviewer that it would be interesting to include more electrodes in the analysis. However, as our aim is to replicate Mathewson et al.'s findings, we will limit the RR to the same analysis conducted in the original study, although we do not exclude that exploratory analysis might be performed. Additionally, we have calculated our sample size on the results reported in the original paper.

It is correct that Oz was included in the paper, but here we have decided to replicate specific results, i.e., the effect of alpha phase on N1 amplitude. As reported in the original paper, the N1 component was measured at Fz. Pz was chosen because it is the electrode at which the alpha phase has been calculated in Mathewson et al., 2009.

Quality checks: Can you please add a bit more description of what "target only trials larger than 0dB…" As the timing is crucial to this study and finding effects (if they are there) then the checks are important. You don't want to find yourself in a situation where a lab fails to replicate due to poor timing, not due to the effect. The relatively high refresh rate will go some way ensure stim presentation is accurate, but of course other things can cause delay. If the results are different across labs then you might want to check with e.g. a photo diode that any results are not due to poor timing. Just to rule that out.

Thanks for this comment. As pointed out in response 6 to Reviewer 1, in the first version of the manuscript, we have not specified that each Lab must check the precise timing of the experiments by using a photodiode before starting the experiment as a requisite in the #EEGManyLab project. This information has been introduced in the present version of the manuscript.