Revision as invited by the recommender, Maxine Sherman.

A multilab investigation into the N2pc as an indicator of attentional selectivity: Direct replication

of Eimer (1996)

Dear Dr. Sherman,

We highly appreciate the careful reviews of our paper, the constructive advice, and the consideration of our manuscript as a registered report. Accordingly, we have carefully considered the comments and attempted to make the most of each one. The point-by-point responses (blue font) to yours and each of the reviewers' concerns (italic font) are provided below. Changes in the manuscript are marked in red font; purely editorial changes are unmarked.

Kind regards,

Martin Constant and Heinrich Liesefeld

(on behalf of all listed co-authors)

Decision Letter: Maxine Sherman

Thank you for submission, and please accept my apologies for the delay in getting back to you. Your Stage One submission has received two high-quality reviews from Dr Clayton Hickey and Dr Reny Baykova, and I have provided a third review, which I have appended below.

All reviewers agree that the submission is excellent, and request only very minor points of clarification. Once these have been completed (or rebutted) then I will go ahead recommend the report.

Dr Baykova suggested that Eimer's third hypothesis in the original 1996 paper is also included as a target for replication here - it is not neccessary for you to do this, but if you decide against could you briefly explain why in the Introduction.

Many thanks for the encouraging feedback! We have added the third result to the effect-size justification as suggested by Dr. Baykova. Regarding the justification for this hypothesis, as far as we can tell, there is no additional a priori reason to assume that the forms N2pc is larger/broader apart from that this is the pattern that Eimer (1996) has observed, so that there is nothing substantial on this point that we could add to the introduction.

Review 1: Maxine Sherman

The authors write:

"The most representative result are the effects of contralaterality in Study 2 (which is the replicated study) for electrode pair OL (corresponding to PO7/8 in the 10-10 system) in the time range 220 – 300 ms for form discrimination $F(1,9) = 57.10$, $p < .001$ and color discrimination $F(1,9) = 17.48$, $p = .002$; thus the smallest of these two F values (17.48) is used to compute the effect size"

Could you clarify what is meant by representative result? Do you mean the result in Experiment 2 that is most representative of the interpretations of the original paper?

We have added the following paragraph to explain our choice (p. 7): "Study 2 is, in a sense, more influential than Study 1, because with only one nontarget item, it provides a stronger test of the main

hypothesis that the N2pc is related to target processing rather than the suppression of surrounding nontargets. The spatiotemporal extent of this effect is most influential as it corresponds most closely to the typical analysis window of the N2pc in subsequent studies"

The authors write that they will report a meta analytic effect size for the original pipeline. Will this also be done for the alternative pipelines? If not, how will the across-labs results from these pipelines be reported, and how (if at all) will they be compared to results from the original pipeline?

Yes, this will also be done for the alternative pipelines. We have updated the manuscript accordingly (p. 8).

Typos

We have corrected the reported typos. Thank you for noticing!

Review Clayton Hickey

The protocol is careful and entirely suitable for reproduction of the target study. I see no need for revision.

Thank you for the positive feedback.

P 2 – 'Eimer's finding does not exclude alternative interpretations… might be a composite reflecting both enhancement of the relevant stimulus and suppression of the irrelevant stimulus…' There is nuance that is missed here. In Hickey et al., 2009, we suggested a couple of alternatives. As described in the RR, we concluded with some confidence that the Pd reflects direct suppression of the neural response to the distractor. We were less sure what computational mechanism the Nt represented. We offered the possibility that it might represent some sort of target enhancement. But we kept open the idea that it might reflect a mechanism of distractor suppression (in addition to the mechanism reflected in Pd). That is, the Nt could reflect activation of cells responsible for the 'silencing' of distractor input into cortex responsible for representation of the target (eg. by inhibiting cells in cortical layers receiving lateral or feedback input). This kind of sheltering mechanism would emerge in tissue that is physically proximal to cortex representing the target, and would emerge contralateral to the target (not the distractor).

This idea was broadly motivated by literature emerging at that time on the role of alpha in vision, and on the possibility of GABA-ergic circuits involved in attentional control. One reason to like the sheltering account is that it easily explains why the N2pc is so sensitive to the proximity, presence, and nature of distractor stimuli. It's also broadly in line with the strong role for distractor suppression that emerges in animal electrophysiology studies of attention. To my knowledge, it remains a valid interpretation.

We agree that this is a highly interesting interpretation of the Nt/N2pc and have updated the corresponding paragraph accordingly (p. 2). We don't go into much detail here, but mention the possibility that the Nt might reflect the suppression of irrelevant stimulation around the target (if there is any such stimulation; p.2).

only few N2pc studies have presented the relevant stimulus without surrounding elements…' We went even a little further than this in Hilimire, Hickey, & Corballis, 2011 , Psychophysiology, and only presented one stim at a time.

We agree that this study is of interest to the readers of the present manuscript and now mention this work in the revised manuscript.

Review Reny Balkova

On page 7 of the manuscript, the authors have included a data sharing statement, stating that they will make raw data and analysis scripts publicly available. Just to confirm, will you also share the data of participants who have been excluded from the analysis?

Our apologies for overlooking this point. Yes, we will share the data of all participants who completed the task but were excluded during data analysis. This is now specified in the manuscript.

On page 8 of the manuscript, the text states that some pilot data has already been collected. Would it be possible to share this data, so that I can run through the appropriate scripts that are on the online repository?

Of course, we have now uploaded the data on the OSF repository. Please note that the analysis pipeline crashes when processing the Essex dataset (it contains bipolar hEOG (i.e., only one channel) which results in a crash; future datasets will contain both hEOGs as separate channels with the same reference as scalp electrodes).

Skeleton scripts for EEG pre-processing, the analysis conducted by individual labs, and the meta-analysis have already been shared on osf and github. Would it be possible to add the github repository in the manuscript as it is a bit easier to interact with the files on github than on osf?

Following your advice, on page 4 we added a link to a Zenodo archive of the repository (https://doi.org/j7kb; which also prominently links to the GitHub on the landing page) to the manuscript. A link to the repository was also added in the author notes. The repository is also a component of the OSF project (which is most likely how you found it).

On page 7 of the manuscript, under "sample size and inclusion criteria", the authors provide a power analysis to justify the sample size that will be collected by each participating lab.
I was wondering whether it would be appropriate to do a power analysis for the meta-analysis as well? Is this something you have looked into?

During our internal review process prior to manuscript submission, we consulted the meta-analysis expert advising the #EEGManyLabs project team. We consulted them again following your comment and they replied:
"A power calculation is not needed, because unless the heterogeneity is really extreme, the overall power will exceed that of any given replication. That said, if you wanted to do a quick power calculation, you could use the approach described in the attached paper (Valentine et al. 2010; https://doi.org/bzk3gg), which just involves picking tau, mu, and an average within-study standard error."
Therefore, we do not think it is necessary to run such an analysis.

When computing the sample size to be collected by each lab, the authors report two of the three numerical results in Eimer (1996) that they are attempting to replicate. For completion, I would suggest reporting the original results regarding the third hypothesis as well (the comparison of N2pc amplitude between forms and colours). It won't change anything in the power analysis itself because the associated F-value is larger than the one used to compute the required sample size, but I think it will be good to have it there.

Thank you for this suggestion; we have updated the paragraph as suggested (p. 7).

Would all participating labs use the same instructions script, and could you share the instructions (apologies if this is already in the experimental files that have been shared, I didn't manage to get to those)?

Yes, so far, all labs have agreed to use the OpenSesame experimental program, which contains the (basic) instructions. If a lab joins and cannot/does not want to use the provided experimental program, we will ensure that their implementation is consistent with ours.

The instructions for letters:

"You will now see letters.

The target letter can be an M or a W.

If the target letter is an M, please press [A KEY].

If the target letter is a W, please press [ANOTHER KEY].

Please respond as quickly and accurately as possible

and maintain your eyes on the central fixation cross."

And for the squares:

"You will now see squares.

The target square can be blue or green.

If the target square is blue, please press [A KEY].

If the target square is green, please press [ANOTHER KEY].

Please respond as quickly and accurately as possible

and maintain your eyes on the central fixation cross."

Between brackets are variables, which are counterbalanced, e.g., [A KEY] can be "left arrow" for participant X and "right arrow" for participant Y.

For completion, could you add the viewing distance that participants will sit at?

This will depend on each lab's configuration and SOP. We will report the viewing distances of each lab when publishing the full report. Please note that the size of the stimuli is specified in dva in the experimental program and that the experimental program does not assume a default viewing distance (i.e., the program crashes on purpose until modified to add the lab's viewing distance), therefore the size of the stimuli should be consistent across labs.

On page 4 of the manuscript, it says that the practice block will run "until the experimenter judges from the HEOG waves that participants are holding their eyes sufficiently still". Does that mean until the absolute magnitude of the HEOG is below 25 μV for a particular amount of time? Does the practice block have a minimum length in terms of numbers of trials to ensure that participants have

been presented with all target configurations? Will the practice block be repeated with participants start the second condition?

This is one area where the experimenter (i.e., person doing the data collection) has a degree of freedom (though not a problematic one, since it happens before data is collected and without knowledge of the outcome). In other words, there is no magnitude threshold for a certain number of trials to end the practice block. This is how the original experiment was conducted (information obtained through correspondence with Martin Eimer) and we felt it should remain this way to stay close to the original study.

The practice block does not have a minimum number of trials. As the task is very easy and the display configurations are equivalent from the perspective of the participant, we do not think that encountering all display configurations during practice is crucial. Having said this, as the number of display configurations is quite low (n = 6 per condition), participants will likely see all potential configurations before the experimenter is convinced that the participant is able to maintain fixation and/or eye movements are reduced to a satisfactory amount.

Yes, the practice block is repeated when the second half of the experiment starts. This is now specified in the manuscript (p. 4).

On page 4 of the manuscript, it states that "more laboratories might join after the in-principle acceptance". Could you justify the decision to include additional labs after the in-principle acceptance? Could you also add more details, for example: (1) how many labs you will be looking to include ; (2) when would be the deadline for additional labs to join – I imagine ideally that would be before data collection starts?

We don't believe there would be any adverse effect of including more labs (since each lab is providing an independent dataset of 28 participants), and we also believe that more labs will mean a more robust replication. Therefore, we do not have a threshold on the number of labs, but we are also satisfied if it remains as is (i.e., if no further lab decides to join).

We do not have a deadline for additional labs to join, the only condition is that the lab's data collection (i.e., 28 participants) should be completed within 1 year after the in-principle acceptance.

On page 4 of the manuscript, it states that "The first "Original" pipeline is the direct replication attempt, and the alternative pipelines will be used to cross-validate the results with more modern processing techniques". What would this cross-validation consist of in the paper?

The cross-validation would mainly consist in checking whether results from all pipelines converge towards the same conclusion. Additionally, we will likely comment on the relative power of the various pipelines to detect the effects of interest, based on the respective (meta-analytical) effect-size estimates.

The first two hypotheses are based on the assumption that N2pc reflects selective attention allocation towards a target stimulus, and I think this is reasonable. The authors mention that these hypotheses are also consistent with other theoretical interpretations of the N2pc (more on this below).

It would be good to add a theoretical justification for the third hypotheses as well – the expected difference in N2pc between forms and colours is only justified. Unless I am missing something, I think

the only justification for this hypothesis is that Eimer (1996) report a difference between the two conditions.

That is indeed the main justification. As far as we can tell, there is no additional a priori reason to assume that the forms N2pc is larger/broader apart from that this is the pattern that Eimer (1996) has observed. We can and likely will speculate on why that is, but before doing so, it would be useful to find out whether the pattern replicates robustly.

*14. Related to this, I think it would be worth explaining why the last column in Table 1A, "**Theory that could be shown wrong by the outcomes**", is filled with NAs. If the study cannot inform theory, I think it would require further justification.*

We believe that it makes much sense to fill this column with N/A for the reasons given in response to your next point and because we have seen this in other registered reports on PCI-RR, for example:

https://osf.io/e8ws2?view_only=c8ec62553146496e8b5e4d100a0f08b5

https://osf.io/bjt37

15. I am not too familiar with the literature around the N2pc so I cannot provide a detailed review on this aspect of the paper. To me, the introduction provided a short but clear overview of the relevant literature and the competing interpretations of the N2pc. Maybe one thing I was left wondering after reading the introduction is why replicating Eimer (1996) is important if it can't distinguish between the different potential interpretation of the N2pc. Could you elaborate on what it would mean for the field if the effect doesn't replicate?

We do not believe replication studies typically prove or disprove a theory. Rather, a non-replication would undermine the support for that theory previously provided by the target study (but does not show it wrong, because it is not designed to, i.e., does not produce data that is incompatible with it). A replication of the present target study cannot be in conflict with any upheld theory, because all current theories on the functional role of the N2pc have taken the classical Eimer (1996) finding into account. A non-replication would cast doubt on the assumption that the N2pc is related to target processing and could push the field to try and test this more directly by designing new studies that might yield results which are obviously incompatible with this assumption (e.g., by employing additional alternative measures of attentional selection).

Furthermore, studies targeted for replication in the #EEGManyLabs project were selected based on their impact, not because they are crucial for any specific theory. Most of us (including the first and last author) joined the project, when this selection process had already taken place. Thus, within the scope of the larger project, a replication/non-replication is informative regarding the replicability of influential EEG studies. This discussion does not seem to belong into the present manuscript, though, but must be part of a more general paper on #EEGManyLabs itself, which will be produced at the conclusion of this project.

Comparing the pre-processing steps in the "Original" pipelines and the pre-processing steps in Eimer (1996), I think Constant et al. have followed the original paper closely. I think all differences in the pre-processing between the two studies have been noted clearly by the authors, and I don't expect

these differences to have an effect on the results. I have a few questions on the statistical analysis. Could you describe how the statistical analysis in the original paper was conducted? From reading the description of the analysis on page 227 in Eimer (1996) and the ERP results on page 230, I can't determine with great confidence what they did. It looks like they computed several ANOVAs – one with all the data, and then two more with the two task conditions separately. Could you also say why you decided to conduct the statistical analysis in the "Original" pipeline differently? Could you also say why you didn't consider analysing the data using a 2X2 ANOVA with experimental task added as a predictor in addition to electrode laterality?

This is also the conclusion we reached as to which tests were run in the original paper.
We use paired-sample *t* tests because they more directly test the effects of interest while being mathematically equivalent to the original ANOVA. This has the benefit of added clarity when reporting the results as well as allowing us to conduct directed tests (which is the correct approach here, since we have clear predictions regarding the direction of the effects). The 2 × 2 ANOVA interaction is tested directly with the *t* test for the third hypothesis.

Finally on the EEG analysis, could you say why you decided to only replicate the results for the occipital electrodes, and exclude the analysis of the parietal electrodes?

The occipital electrodes (OL/OR) are today's PO7/PO8 which are the electrodes from which virtually all labs measure the N2pc (either by analyzing only data from these two electrodes or from lateral clusters centered on these electrodes). This makes the analysis of Eimer (1996)'s "occipital electrodes" the most relevant/influential analysis. This electrode pair is the spatial peak of the component of interest, and analyzing only these allows the paper to be more focused.

Are you planning on replicating the behavioural analysis in the original paper?

No, the behavioral results are of no relevance for the actual research question. Should an unexpected behavioral result arise, we will consider reporting it as an exploratory result.

As with many neuroimaging studies, the results would be confounded by some amount of motor preparation. I think motor preparation would be of concern in this task also because the response keys – left and right – correspond to the different targets rather than the target location. Could you discuss this? Also, would responses be given with different fingers in one hand (e.g. right hand only, index finger presses the left button and middle finger presses the button) or with two hands (e.g. index finger of left hand presses left button, index finger of right hand presses right button).

Participants will respond with their left or right hand as in the original study. Indeed, we envisage that most labs will ask participants to use their index fingers to give this response. However, we don't readily see how motor preparation could confound the present results. Participants will respond to half of the targets on the left side with their left hand and to the other half with their right hand and the same holds for targets on the right side, Thus, target side and response hand are deconfounded. Moreover, at the analysis stage, targets are aggregated (i.e., "blue" and "green" become "color", "M" and "W" become "form"). Unspecific motor preparation (e.g., in anticipation of the display onset) also does not play a role here, because we analyze the difference between electrodes contra- vs. ipsilateral to the target. That any such non-specific activity is subtracted out is a major advantage of the lateralization method underlying the extraction of the N2pc. In a way, activity measured at ipsilateral electrode sites acts as a control for any unspecific activity.

Previously, we had not defined a clear and strict modality of response, but because this comment has fortunately made us aware of this point, we decided to standardize this to using both hands, as in the original study.

Have you considered the possibility of using an eye-tracker to confirm that participants remain fixated at the central cross?

Yes, we have considered this potential requirement but we decided against it because no eye tracker was used in the original study and not all labs have access to an eye-tracker. Asking only those labs that routinely employ an eye tracker to use it would add more across-lab variation to the experimental procedure, not only in terms of the degree of control over fixations, but also via side effects (i.e., use of chin-rest [which can also create muscle tension], need for calibration and re-calibrations [increases experiment duration and participants tiredness]). In fact, the fixation control (that could be achieved by an eye tracker as suggested) is in this research tradition often implemented by monitoring the EOG (as will be done here).

What is the reason for doing all 6 blocks associated with 1 condition followed by the 6 blocks associated with the other condition, rather than intermixing the order of the blocks?

We replicate the original experimental procedure here. This was probably done in order not to confuse the participants, but rather to let them build a firm attentional set (i.e., a mental representation of the task, a target template mainly) rather than switching between attentional sets. We now emphasize that also this aspect of the experimental design replicates the original study (p. 4).

Before going through the repository, I was going to suggest including Bayesian t-tests for each of NHST t-tests. After going through the repository, I saw that the paired t-test script is set up to compute Bayes factors using the JZS prior, but if I am reading the code correctly it seems you have currently decided not to calculate Bayes factors (there are also not listed in the manuscript amongst the statistics you are planning to report). Therefore, I am going to proceed with my recommendation to include Bayes factors and to compute them using the approach detailed in Dienes (2014). Since your hypotheses are directional, my suggestion is to define the prior as a half-normal distribution with a mode of 0 and a standard deviation equal to half the prior-expected effect size (the prior effect being the difference in amplitude found by Eimer). think this would be very useful in case some of the effects don't replicate because you will be able to see if you can provide evidence in favour of the null.

The paper on Bayesian analysis I reference is available here and it details the approach I described above: Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. Frontiers in psychology, 5, 781.

Thank you for this suggestion. As we were not familiar with the procedure by Dienes (2014), we have looked into it carefully. However, its potential advantages compared to the "standard" BF *t* tests (as described by Rouder) that we are more experienced with were not evident to us. Nonetheless, we are convinced that reporting BF *t* tests is generally useful and we will perform these tests in JASP (our current implementation of BF *t* tests in MATLAB is only for two-tailed *t* tests). This addition is now reported in the methods section (p. 6).

In Table 1A, could you add the refresh rate of the monitor that will be used by the lab in Málaga, and the resolutions of all the monitors?

We have added this information to the table.

What is the reasoning for the differences in the impedance thresholds between labs? And aren't these thresholds a bit high?

Each lab has its own SOP; the idea of #EEGManyLabs is to not force labs too far away from their SOPs. All impedances are in any case equal or lower than that recommended by the manufacturer of the amplifiers. But it certainly is an interesting follow-up question to re-analyze the data from this and other #EEGManyLabs studies with regard to the impact differences in impedances across labs have on the quality of the results (although, the lab climate might play a larger role; see Kappenman & Luck, 2010; https://doi.org/bz8dxg). These are the types of questions that will be explored by the final paper in this project.

At the top of page 7, the last bullet point above "sample size and inclusion criteria" references with: "… forms or letters…" . Did you mean to say "forms or colours"?

On page 7, under "sample size and inclusion criteria", the text says: "The most representative result are the effects of contralaterality in Study 2 (which is the replicated study) for electrode pair OL (corresponding to PO7/8 in the 10-10 system)". Did you mean to say "electrode pair OL-OR", or am I misinterpreting something?

Yes to both- thank you for noticing!

Can you explain the existence of the target-only trials if they are excluded from the analysis?

Eimer used these trials for some of the reaction times analyses. We include them to stay as close as possible to the original procedure (but they cannot produce an N2pc since targets are presented in both hemifields).

The next comment is entirely subjective, so feel very free to ignore it. If I was making Figure 1, I would extend the shaded area which represents the time window of interest to cover the full height of the graph. There is something about the off-centred placement of the time window indicator in 1A that I find very unnerving.

Thank you for this suggestion. We agree that this looks neater and have updated Figure 1 accordingly.