Utrecht University, Faculty of Humanities

Department of Languages, Literature and Communication

Trans 10, 3512 JK Utrecht, The Netherlands

Telephone
+31 30 253 7221

E-mail
i.m.vanderwulp@uu.nl

To:
Peer Community in Registered Reports
Dr. Elizabeth Wonnacott

Date
September 27, 2023

Subject
Stage 1 submission of revised manuscript

Dear Dr. Wonnacott, dear Reviewers,

Thank you all very much for these very insightful and helpful reviews on our manuscript. We have now revised it and hereby resubmit the manuscript to you. In this document, you will read a point-by-point reply to your comments and questions stating what changes we made to the manuscript. Attached in the OSF folder you will find the revised version of the manuscript, as well as two supplements which are new in comparison with our previous submission. We have highlighted the changes in the manuscript.

Dr. Elizabeth Wonnacott:

1) *Thank you for submitting your work as a registered report to PCI RR. Your paper has received two reviews from experts in the area and I have also read the paper.*

   Both reviewers are positive, though they raise some important points about design and analyses. I agree that the paper is well written and that it potentially taps into some interesting questions. However, I concur with the reviewer concerns, and I would particularly like to emphasize and enlarge on the point made by the first reviewer about the difficulty in keeping track of your different questions in the analyses plan and – critically- the fact that your sample plan is not taking into account all of the different questions you are answering.

   A requirement for being accepted as a Stage 1 RR is that for every RQ/hypothesis you are testing, you have established that you will have a sufficient sample. Every row in the big table at the end has to have its own sample size justification (NB a member of the managing board confirmed for me that this is the case). For places where you don't want/ aren't able to do that, you can keep those as exploratory analyses that you look at once you have the data, but then they aren't part of your pre-registered analysis plan. This means they can't feature in the abstract and – though you can talk about them in the discussion - they can't be the main focus.

   At present, you only seem to be powering for RQ1, which is a main effect. But your other RQ's will have different power requirements. You need to find a way to estimate for every hypothesis you want to retain. I think that this will most likely require simulating random data.

   I also would like to comment specifically on the Bayes Factors, first I think it's great to use a method that can allow us to distinguish between evidence for H0 and ambiguous evidence, especially for the kinds of questions you are interested in (as you note, there are many papers with null effects in the literature looking at relationships between statistical learning and other factors, and it is hard to know how to evaluate without an inferential statistic that evaluates evidence for the null

- Bayes Factors can avoid this problem). It would be better to be consistent be consistent in your approach though, so if using Bayes Factors for some tests, use throughout. (Or at least if you think there is there is a good reason to use in some cases and not others, this needs clear justification). Note that for Bayes Factors, when you do your sample size simulations you need to separately consider sample required for establishing H1 when H1 is true and also H0 when H0 is true (i.e. for the later, you will need to test simulated data sets where you set the parameter for the effect of interest to 0). A couple of points here: it will be much harder to find evidence for H0 being true than for H1 being true. You are very likely not going to be able to plan for 90% power of getting a chance of BF<1/6 when H0 is true. So its going to be about balancing the ideal against feasibility, but you need to show that you don't have (e.g.) only a 20% chance of finding evidence for H0. I think it would be reasonable to use a criterion of BF<1/3 rather than BF<1/6, but this will still be challenging. I note that you plan to use the factional Bayes Factor approach- I can see the appeal as it seems to avoid having coming up with value for you predicted effect which you need for an informed model of H1. However, this may make it very hard to find evidence for H0 when it is true. You can ascertain this through simulation. An alternative is to use an informed model of H1, which uses an estimate of predicted effect size. Although I know that in some cases you don't have values from previous work to use, there are heuristic you can apply. I strongly recommend this paper by managing board member Zoltan Dienes: Dienes Z. How Do I Know What My Theory Predicts? Advances in Methods and Practices in Psychological Science. 2019;2(4):364-377. doi:10.1177/2515245919876960) (the ratio of means heuristic might prove useful?).

While coming up with predicted effect sizes may seem challenging, note that you will in any case need some rough predicted effect size estimations in order to do power calculations (i.e. whether for frequentist or Bayesian analyses, and also whether you use power software or do it via simulation).

(NB- if you would like help with simulating data for mixed effect models I could share some scripts with you to help get started- feel free to reach out separately by email. You might also find this preprint https://psyarxiv.com/m4hju on which I am an author useful for some practical demonstrations of how you can combine linear/logistic mixed effect models run using the lmer package and Bayes Factors. An alternative – and more commonly used - is that you can use the brms package and use model comparison (this is also briefly discussed at the end of our preprint). Both approaches require you to have an estimation of your predicted effect to inform model of H1).

Thank you very much for your assessment of our manuscript, and your additions to the reviewers' comments. We had indeed estimated power for the main effect of the replication of Batterink & Paller (2017), but we had not provided sample size argumentation for the other research questions. Using data simulations based on LMMs – Elizabeth, thank you so much for the scripts you sent us to do this! – as well as Bayes Factor Design Analysis for correlations, we came up with sample size justifications for all research questions. These are now incorporated in section 2.1 of the manuscript (pp. 12-13), as well as in the study design table (appendix A). We agree with you that a Bayesian approach throughout is preferred, as this would be more consistent instead of doing this for the correlations only. We have chosen to go with this approach, such that all proposed analyses now perform Bayesian hypothesis testing. An additional advantage of Bayesian hypothesis testing is the possibility of Bayesian Updating, which is the approach we want to take for our sampling plan. We established a maximum feasible sample size of 105 participants for this study, based on time and funding limitations. This includes an initial sample of 45 participants

(identical to Batterink & Paller, 2017) and four updates of 15 participants per update. As you will see in the study design table (appendix A), we are in most cases indeed not able to plan for 90% power when H0 is true, adhering for BF_{01} < 1/6. We have taken your advice into account and calculated power for BF_{01} < 1/3 additionally. We are indeed balancing the ideal against feasibility, as you wrote earlier. All results of our simulations are included in the study design table and the supplements with the code output. We believe that with our updating approach, a maximum sample size of more than 100 participants, and – for some of the research questions – a robust line of earlier research, we will have a reasonable chance of answering our research questions.

Importantly, a student who recently wrote her thesis with us collected pilot data (N = 15) that proved useful input for some of the aforementioned sample size estimations. Her thesis considered a different question than the study under review here, where the structured condition was her control condition (named the isochronous condition in her project, see www.doi.org/10.17605/OSF.IO/MA2C6 for details on her project). She had additionally collected but not analyzed the tests for individual differences we propose here. This was also valuable input for some of the sample size estimations, as we could use this pilot data to estimate some of the possible effect sizes. We will not use any of this pilot data in the analyses for the stage 2 report.

With regard to priors, where available, we used effect sizes from earlier published research and our behavioral pilot study that was already present in the manuscript (appendix B) to establish an informed model of H1. With regard to the correlations, we will adhere to the default prior κ = 1. In two cases (the rating task and the mediation analysis) we will use the fractional Bayesian approach with Bain. In those cases where we do not have an informed model of H1, we will follow the initial analysis with a sensitivity analysis. We have used our student pilot for the sample size estimation of the mediation analysis, but believe it is more conservative not to go with this pilot data as a model of H1 for the final analysis (due to the very small N and added condition). Finally, we performed Bayes Factor Design Analysis for a range of possible effect sizes as well as H0, which paints a clear picture that we will have a reasonable chance of finding an effect if it exists with our proposed sampling plan.

Furthermore, motivated by the comments regarding the sample size estimations, we changed the position-controlled baseline condition into a truly random condition in order to stay as close to replicating Batterink & Paller (2017) as possible. We realized that the baseline condition proposed in the earlier version will not be comparable either to Pinto et al., (2022) or Batterink & Paller (2017). Regarding Pinto and colleagues, they used a structured condition with 6 words and based their baseline condition thus on a larger number of syllables per position. With only 4 words in our stimuli (as in Batterink & Paller, 2017), we predict that we might not find the same effect of condition and given the sample size discussion raised by the reviews, we would not have any comparable experiments on which to base the sample size estimations and models of H1. We were able to make sample size estimations by deciding to go for a closer replication of Batterink & Paller (2017) and thus to use a random condition instead. We will use the syllables from set B from our pilot experiment for this and have changed this in the manuscript and study design table accordingly.

2) *One other thought with regard to power: I wasn't clear why it was necessary to do a bimodal split on the SSS task, rather than treating as a continuous variable. Might not the later been more sensitive?*

Thank you for this suggestion. We decided that we will indeed treat the PLV of the SSS data as a continuous variable. This will also address the questions asked by Reviewer 1 about how many participants will be in each group and how the cut-off between high and low synchronizers is determined, since there will no longer be a group distinction.

3) *So in sum, I think this is potentially a very interesting a paper, and I would love to see it as an RR at PCRI. However, you will need to find a way to show that you have resources to get an N that has a reasonable chance of giving evidence for H1 if it is true and – if using Bayes Factors– for H0 if H0 is true.*

Thank you for your kind words about our submission. We hope that we have now convinced you with this revised version of the manuscript, that we will have a reasonable and sufficient N to be able to test our hypotheses.

Reviewer 1:

Thank you for the opportunity to review this manuscript. I hope that my comments will be helpful for the authors to revise their experiment plan and write-up.

1) *The introduction was clearly written. I also believe that the rationale of the proposed hypotheses is clear. I like the idea of testing rhythmic/musical abilities as an underlying mechanism of statistical learning ability, and believe that the scientific questions were overall well-motivated.*

Thank you very much for your positive assessment of our manuscript and study.

2) *The research discussed reminded me of the "auditory scaffolding hypothesis", which proposes that deaf and hard-of-hearing children/adults may have worse sequential statistical learning due to less access to auditory signals (though evidence for this hypothesis has been lacking in recent years). I wonder if it is worth integrating it in the introduction given it postulates a similar mechanism as the manuscript here (auditory processing impacting more general statistical learning)?*

Thank you for this insight. We understand how our manuscript could remind you of the auditory scaffolding hypothesis. However, our hypothesis of rhythm processing influencing auditory statistical learning, is a domain-specific proposal that does not make predictions about SL in other modalities. This remains to be further researched and is out of the scope of our paper. Moreover, the auditory scaffolding hypothesis is indeed faced with counterevidence from multiple studies finding that deaf and hard of hearing individuals are indeed capable of SL, and do not show a delay or deficit when compared to a normally hearing population (e.g., Terhune-Cotter et al., 2021).

Terhune-Cotter, B. P., Conway, C. M., & Dye, M. W. G. (2021). Visual sequence repetition learning is not impaired in signing DHH children. *The Journal of Deaf Studies and Deaf Education, 26*(3), 322–335. https://doi.org/10.1093/deafed/enab007

3) *There are a lot of different predictions at the end of the introduction, and I found it hard to keep track of all of them (and the different associated tasks). Is there a way to visually present the different components of the experiment and their predictions? A simplified version of the table that is provided at the end of the experiment may also work for this purpose. (Relatedly, I wonder*

if this paper is doing a bit too much by trying to investigate rhythmic/musical abilities' contributions to statistical learning, but then also looking at working memory and vocabulary size via exploratory analyses – it might be easier to focus the manuscript if the rhythmic/musical abilities were the clear focus.)

Thank you for this suggestion. We have now added a figure in the introduction (Figure 1, p. 10), aimed at clarifying the different predictions.

It is true that we are aiming to investigate quite a lot of components that could underly individual differences in SL at once. Our hypotheses regarding rhythmic abilities are of particular interest. However, working memory and vocabulary size may also potentially be related and necessary pieces of the puzzle. Tasks aiming at measuring rhythmic ability such as the PROMS and CA-BAT require the participant to listen to multiple stimuli and compare them. Working memory is a key component of this process. Consistent with this idea, we actually found in our thesis student's pilot sample (N = 15) that these rhythmic ability tasks and the digit span correlated positively (e.g. PROMS: BF_{10} = 9). Furthermore, theoretically, one product of SL in the domain of speech segmentation is vocabulary, but this link has previously been researched only in children. So yes, we are aiming to investigate a lot, but each of these components represent part of the bigger puzzle we are aiming to understand.

Method:

Please be more precise when describing some of the methodological decisions:

4) *Please justify power in more detail for the different DVs.*

Thank you for pointing this out. We have now done this and added the information in section 2.1 (pp. 12-13), as well as in the study design table (appendix A). We estimated power through data simulations, Bayes Factor Design analysis, and references to previous research. Output of the simulations can be found in the supplementary materials. See also our more elaborate reply to Dr. Wonnacott in point 1 above.

5) *What will count as "approximately 35 participants in each of these groups" (p. 12)? Does that mean it is ok if there are 40 people in Group 1 and 30 in Group 2? What about 45 and 25? How is this decision made?*
   - *How exactly is the cut-off between high and low synchronizers determined? Why is this distinction needed in the first place and would an analysis with synchronizing ability as a continuous variable not be more appropriate anyway?*

We have decided that we will indeed treat the PLV for the SSS task as a continuous variable. Therefore, there will no longer be a group distinction in our participant sample. See also our reply to Dr. Wonnacott's second point above, which is congruent with yours.

6) *What exact criteria will be used to identify bad channels? (p. 16: "Bad channels identified upon visual inspection of the data or during data collection will be interpolated")?*

During data collection, the bad channels will be identified by their impedance, which we will keep below 20mV. We noticed that this was indeed not yet included in the manuscript and have added

it to the revised version in section 2.4 (p. 17). During analyses, bad channels are identified visually, if they show frequent large drifts or noise, this information is also added in section 2.4.

7) *What are participant exclusion criteria? Please list these clearly. For example, it is mentioned that "data of participants can be excluded after participation in the case of technical issues that cause a premature termination of the experiment, or for failure to comply with instructions in general" (p. 12). Is it possible to pre-specify exact exclusion rules here (e.g., participants will not be included unless they followed instructions on X trials of task X)?*

   We have replaced *failure to comply with instructions in general* with the following more precise criteria in the revised manuscript (section 2.1, p. 13):

   - If the participant has < 50% targets detected in the target detection task. In our pilot experiments and earlier studies from Laura Batterink, all participants were above this percentage.
   - If the participant wishes to retract/stop their participation during the experiment.

8) *What kind of data quality checks will be used?*
   - *Will you check for hearing issues? Anecdotally, I have heard that a non-substantial percentage of students actually do not hear normally at all frequencies, but may not necessarily be aware of this. If a participant had (unknown) hearing issues, how could this affect your results?*

   We instruct possible participants that they cannot participate if they have any history of hearing impairments, as well as tinnitus specifically (we have now added this specifically to the manuscript as well). Since we present speech stimuli at a fundamental frequency of 100 Hz, we think this is sufficient to ensure participants can perceive the stimuli well enough, as would also be confirmed by their performance on the target detection task (see point 6 above). The results of the pilot experiment also indicate that self-reported hearing ability is sufficient to ensure adequate hearing for our purposes. With regard to the rhythmic tests, it is important that participants can perceive the rhythm of the music, but less emphasis is placed on perception of the tones.

9) *Why do you plan to use a forward digit span and not a backward digit span as a measure of working memory?*

   We chose to use the forward digit span and not the backward digit span because the forward digit span is associated with verbal working memory and depends on the phonological loop, which is the most interesting for our study. The backward digit span, however, is more so associated with executive functioning and cognitive control. This is now briefly described on p. 11 in section 1.6.

Other minor comments:

10) *p. 4: "foils that were not present" – this could be easily misunderstood as foils being completely new (instead of syllable combinations that had not been present); please reword.*

   Thank you – we have reworded this as "syllables presented in a recombined order " (p. 4).

11) *p. 17: I do not understand the difference between the entire exposure period and time course of exposure ("will then be calculated for each participant over the entire exposure period, as well as over the time course of exposure"). Please reword.*

We have reworded this as follows (bolds not included in the manuscript, but indicate the changes made): "The Word Learning Index (WLI) will then be calculated **as a mean** for each participant (computed over the entire exposure period), as well as **for each epoch bundle** over the time course of exposure, for both the structured and random conditions." (p. 18).

12) *p. 20: How exactly are values standardized?*

Thank you for pointing out this lack of specification. The values will be standardized by subtracting the mean from the variable, and subsequently dividing that by the standard deviation of the variable. We have added this information in the manuscript (section 2.5.3, p. 22).

13) *Not all references are in APA format (words within article titles should not be capitalized).*

Thank you for noticing this. We have now corrected the references.

14) *Consider moving the information about the pilot experiment (section 3) into a supplement. As mentioned previously, the experiment and all the resulting hypotheses are already very complex, thus making it hard to follow along exactly what was done – going back to the pilot after reading about the experiment feels disjointed.*

Thank you for the suggestion. We have moved the pilot experiment from section 3 to appendix B.

15) *Consider changing the title to "Investigating individual differences in auditory statistical learning and their relation to rhythmic and cognitive abilities", given that it is unlikely that rhythmic abilities contribute to visual statistical learning (correct?).*

We have changed the title of the manuscript upon your suggestion. However, we used the word *linguistic* instead of *auditory,* to be as specific as possible that we are looking at language and not tones.

Reviewer 2:

This basically looks fine to me - it's a relatively complicated behavioural and neural study looking at relationships between musical experience/ability (possibly reflecting more basic aptitude for rhythm) and statistical learning, spacifically statistical word segmentation. The hypotheses are well-motivated by the literature; the experimental protocol is quite complex but explained in sufficient detail. I did however have some concerns about the planned statistical analyses, and two question about the design. My concerns are:

1) *One of the main behavioural measures of statistical learning comes from the 4-point familiarity rating that participants give to words/part words from the speech stream. However, it looks like those 4-point ratings will be analysed using a vanilla linear regression, i.e. assuming that response values from -infinity to +infinity are possible. I don't think analysing ordinal response data in this way is best practice and weird stuff can happen (e.g. the model predicting impossible*

*negative values or values off the scale - indeed there will always be some probability mass on these impossible values in this model). I would suggest the authors either revert to the 2AFC method used in the pilot (in which case a logistic regression is fine since it is designed for 2-point response scales and never predicts impossible values) *or* if they want to continue with the 4-point response scale, it is reasonably straightforward to do an ordinal regression in R with mixed effects (I think the package ordinal has a function clmm that will do this for you) which doesn;t have the problem of vanilla linear regression on ordinal data.*

Thank you for your positive assessment of our manuscript and study, and for expressing your concerns. We note your very valid point that the previously proposed LMM approach is not suitable for these ordinal data of the rating task and thank you for the suggestion of analyzing this data with the *clmm* function from the package ordinal. We changed the manuscript (section 2.5.1, p. 19) and study design table (appendix A) to accommodate that approach.

We would like to keep the 4-point rating instead of the 2AFC, as this allows us to stay close to the original study of Batterink & Paller (2017), and because the 4-point scale may also be more sensitive than the 2AFC. The rating task also has certain advantages over the 2AFC (e.g., it does not test the same words multiple times, which can produce memory interference over the course of the test).

2) *You should include by-participant random slopes for factors which vary within-subjects - e.g. since condition (structured vs baseline) is within-subjects, any analysis which includes condition should have a random effect structure at least as complex as (1 + condition | participant). Otherwise the model is forced to assume that all participants show the same effects of condition, which can be anti-conservative (e.g. if one or two participants show a huge effect that will drag the effect of condition up in a model with no random slope; in a model with random slopes it can handle these outliers without messing up the overall estimate of the effect of condition).*

Thank you for this suggestion. We will add condition as a random slope indeed. We have already incorporated this approach in the simulations for sample size estimation as well (see supplements). In the manuscript this information is altered in section 2.4.1. (p. 18)

3) *It's not super-relevant, but in case you end up using a 2AFC approach in the main experiment: in your pilot experiment, you don't need to (and indeed shouldn't) run a t-test on %s to check if participants are above chance (same reasons as given above - t-tests don't know that %s are bounded by 0 and 100). Your logistic regression can tell you if participants are on average above chance if you look at the intercept (and code condition appropriately, e.g. using sum contrasts so that the intercept reflects the grand mean) - any intercept significantly above 0 indicates above-chance performance, since the intercept is the log-odds of a correct response and log odds of 0 = odds of 1:1, i.e. 50-50 responding.*

Thank you for pointing this out. We will not use the 2AFC in the main experiment (see point 1). We used the t-test in the pilot experiment only to illustrate the above chance performance that is consistent with previous research. Either way, our participants in the pilot scored significantly above chance.

4) *Design question: Given that the structured sequence learning task always comes first, how do you rule out the alternative explanation that the difference you see between conditions in your*

statistical learning task is partly driven by order? Or does that confound not actually matter, given that you are interested in individual differences.

We have extensively debated whether or not to counterbalance the order of these conditions. However, given our focus on individual differences, applying different orders would add another factor to the design, requiring a larger sample size. Therefore, we decided to keep the fixed order. Moreover, it might be the case that there is a spill-over the other way around when the random condition is presented first. In that case, participants could unconsciously remember that there is nothing to learn in these stimuli and learning may be reduced in the second condition. Finally and most importantly, the random condition is there as our 'sanity check.' We are primarily interested in the individual differences that emerge during learning of the structured stream. Therefore, prioritizing our focus on individual differences, we decided to keep the structured stream as the first exposure condition, while participants are completely naïve to the kind of stimuli they are about to hear.

5) *Design question: In the pilot it was a bit of an issue that you didn't have sufficient variability in musical expertise. Are you going to sample differently to avoid this in the main experiment, or are you hoping that a larger sample will naturally uncover more musical individuals? The worry is that if their underlying distribution of musical ability is quite tight, a bigger sample will still have low variance here, which might mean you have quite low power to spot effects of musical ability.*

   We will aim to find more individuals with musical expertise for this sample than we did for the pilot experiment. For the pilot, we only recruited from the participant database of the Institute for Language Sciences, consisting mainly of undergraduate students. We will aim to recruit more broadly for this sample. For instance, Utrecht is home to a conservatory (Utrechts Conservatorium, HKU), where we plan to recruit participants for our experiment.

We would like to thank you again for your reviews of the previous version of the manuscript, and we hope you will be pleased reading the revised version and the responses to your comments and questions provided above.

With kind regards, also on behalf of Marijn Stuiksma, Laura Batterink, and Frank Wijnen, Iris van der Wulp