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
November 30, 2023

Subject
Stage 1 submission of revised manuscript

Dear Dr. Wonnacott,

Thank you for your insightful feedback on our revised manuscript. Your comments were valuable in shaping the initial revision, and we appreciate your continued engagement. We have carefully addressed each point raised and hereby submit the revised version for your evaluation. In this document, you will again 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 a new render of the simulations file. We have highlighted the new changes in the manuscript. We hope that you find the manuscript fit for acceptance with these changes implemented.

Dr. Elizabeth Wonnacott:

- *Thank you for resubmitting your Stage 1 paper. I appreciate the very large amount of work that has gone into this revision. The two reviewers are happy with your changes, bar two very minor suggestions from one of the reviewers which you can easily implement.*
  I also think the paper is much improved. My only remaining concerns are in terms of the analysis plan and sample size simulations. I have spent some considerable time going through your analyses plan and script, and there are points where I don't follow and/or where I have concerns and questions. I have listed these analysis by analysis below. Two key points I want to highlight are:

  (1) I question the choice (for some analyses) to use a default model representing the predictions under H1 rather than one with parameters informed by values that come out of the pilot /previous work. You justify this in terms of being more "conservative", but in Bayes Factor terms, the use of defaults is problematic because it is likely to lead to you to finding evidence for H0 when it isn't true (default priors generally have parameters which assume largely effects are more likely, so if the true effect is small, the data may well be more likely the null than under this H1). (See Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas—part II. Linguistics and Language Compass, 10 , 591–613. https://doi.org/10.1111/lnc3.12207 and Dienes, Z. (2019). How do I know what my theory predicts? Advances in Methods and Practices in Psychological Science, 2, 364-377- and references cited therein on this point).

  (2) For the analyses using Bain, you need to give more details of the model of H1 and you need to have some kind of simulation that demonstrates that you will have a decent both chance of evidence for H1 if H1 is true and (critically) also for H0 if H0 is true. (NB- I previously consulted with Zoltan Dienes for his views on this type of analyses - he pointed out that a potential concern with this method is that you might not be able to find evidence for H0 if it is true, so you need to demonstrate that this isn't the case).

  I would like to ask you to address the points below in a revision. I am aware that you are under time pressure and I apologise that it has taken as long as it has to get to this stage. Note that I won't be sending this back out to the original reviewers, however I can't rule out that I may need to bring in

someone specifically for statistical advice if I can't understand your revision / feel that evaluating it is beyond my expertise. In either case, I will do my best to turn this round for you as swiftly as I can. To speed things up, you can feel free to reach out to me personally if you have questions about the below.

Thank you and both reviewers for recognizing the work that was put into the first revision of this manuscript. We are very glad to read that you all agree it has much improved. We understand that you had questions and comments about the choice for default priors in JASP and about the package Bain. We have now switched to informed priors for the correlations in JASP (1), and added a substantiation on how Bain calculates the Bayes Factor, and prior and posterior distributions (2). We will respond point-by-point to your comments below.

COMMENTS ON THE DIFFERENT ANALYSES

- *Note that in the below I am assuming that all of the tests you mention are to be considered key tests of your hypotheses, and thus we need to know that, in each case, you are planning a (max) sample which give you a reasonable chance of finding evidence for H1 when H1 is true and for H0 when H0 is true. However, some may not be considered tests of critical hypotheses. For example, to replicate the key EEG results, the interaction between epoch and language condition is interesting to test for but might not be key in the way the main effect is? For such hypotheses the sample size analyses are less crucial, but then that affects what you can highlight in the abstract and focus on in your discussion. You need to clearly differentiate in that case.*

  Thank you for pointing this out. We now differentiate between our critical hypotheses and other, more exploratory hypotheses. In Appendix A, we have highlighted in green which tests would be critical for our sample size determination. These are also explicitly mentioned in section 2.1. of the manuscript. Following your suggestion, we have decided that for the main EEG results the difference between the language conditions is critical, but the interaction with epoch number is not. We have altered this in the manuscript. We have also gone over the manuscript to make sure we differentiated which of these tests were mentioned as part of the abstract and main hypotheses of this paper. We hope the distinction is clearer now.

Analysis for the replication of EEG results

- *General: Its great that you have gone full Bayesian, but the reporting needs to reflect this. So for example in Appendix A RQ1 column 6 you talk about interpretation given "main effect of condition" and "no main effect of condition"- you should talk about this in terms of specific Bayes factor thresholds which will leave to you infer that there is evidence for H1 and (critically) also for H0. This should also be stated in the main text (also note that you shouldn't talk about "testing significance" in the context of Bayes Factors- as you currently do on page 18). In addition, when discussing inference threshold in the main text, I would personally also make the point that, though you have thresholds for inference, since BFs are continuous you can additionally interpret them continuously (unlike p values) so the higher the BF the stronger more evidence for H1, the lower the stronger the evidence for H0. (I won't make this point again, but in general for all of the analyses below, also go through the paper and table and remove mention of "significance", and make sure that you are talking equally about the conclusions you will draw if you meet your criteria for concluding evidence for H1, and if you meet your criteria for concluding evidence for H0)*

  Thank you for your comment on the reporting style. We have now altered any mention of "significance" or similar terms, and added the point that BFs are continuous on page 18, along with some useful references. Moreover, we have added discussions of evidence for H0 in Appendix A.

lmer models:

- *in the model looking for a main effect you will aim for (1+WLI|participant) which is a full random slope structure (intercepts, slope and correlation between the two); in the model also including interaction a full random slope structure would be (1+languageC*epochbundle|participant) as both WLI and epochbundle are between participants. Can you justify the choice not to have this structure? (Perhaps your simulations suggest too many convergence errors?)*

  The simulations indeed provided us with many convergence errors – mostly in the form of singular fits – when we adhered to this. We consulted our statisticians again and they informed us that singular fits are unreliable and thus cannot be interpreted. It means that the model is too complicated and overfitted on the data. Related to your comment below on how the BFs compare when the model does or does not converge (under *simulations*), we have altered the supplement such that it splits the output for every lmer simulation in non-singular and singular fits, and computes BFs for both. This way, you can compare the outputs. We only used the non-singular fits as the numbers we rely upon for the sample size determination and are thus reported in Appendix A. Singular fits decrease with an increasing N in the simulations. However, with regard to the full random slope structure, this was not the case. We have now also included these simulations in the supplementary document (under the simulation for the model with the interaction). We had taken them out before because we only wanted to include analyses we will actually implement, but we do understand it is helpful for the reader to see what we tried that did not work and why.

- *You say you will remove random slopes if you get convergence/singularity error- a small point but you could first try removing correlations between slopes – the syntax is (1+ languageC ||participant) - then if that still doesn't converge remove the slope completely (1|participant)*

  Thank you for this input; we were unaware of this possibility. We have now implemented this as an intermediate step for the analysis in case of convergence/singularity errors in Appendix A for RQ1 as well as in the manuscript (page 19).

Bayes factor

- *The value that you set as the sd of the half normal should be a rough estimate of your expected difference between conditions. I am therefore not sure why you take the value from the previous study and divide by 2? Is it becaue you consider this to be a maximum value? Provide some explanation.*

  Our apologies for the misunderstanding. In our simulations we did not divide this by two, but there was some confusion if we needed to do this and this was included in the manuscript by mistake. We have now corrected this in the manuscript and Appendix A.

- *Diennes 2020 (Dienes, Z. (2020, April 3). How to use and report Bayesian hypothesis tests. https://doi.org/10.31234/osf.io/bua5n; see also Silvey et al for a worked example) also recommends reporting robustness regions for every Bayes factor you compute. Robustness regions show the range of values of H1 you could have used and still got the same qualitative result (i.e. Bayes factor within the same thresholds as the one you report) and are a type of sensitivity analysis. (NB – this comment applies for every BF you compute, unless there is some other type of sensitivity analysis you prefer).*

  Thank you for this comment. We have now also included information about the theoretically plausible ranges of values we will use to calculate the robustness regions for the analyses in the manuscript and Appendix A.

Simulations

- *In the table/text under the table can you more clearly say - for the sample you have finally decided to use i.e. N=105- in what % of runs you meet your threshold for inferring evidence for H1 when H1 is true*

(i.e. BF>6) and for inferring evidence for H0 when H0 is true, at the different thresholds of interest (BF<1/3 an BF<1/6). (NB – its important that you have explored the consequences of a higher/lower sample in your simulations- and of different thresholds- but the information in bold is what is needed to assess sensitivity of the planned sample size and this should be quick and easy for the reader to look at.)

Thank you for this comment. We initially performed all our simulations on sample sizes 20, 50, 100, and 150. From there we looked at what was feasible and what the updating sample would be. The samples 45 and 105 were indeed not included. An intermediate step of two updates would be 75. We have now performed all simulations on 20, 45, 75, 105, and 150 participants. The 150 is an infeasible sample, included to indicate that even if we would be able to go that far, we would often not have much more of a chance of finding evidence for H1/H0. However, with 45 and 105 included, we can now show the numbers for our minimal and maximum sample size in a clear way. We have altered this in Appendix A.

- *You have quite a lot of convergence errors/singularity errors returned. Can you maybe show that BF's are roughly the same for the set of runs where the model did/didn't converge?*

  As we mentioned in response to your comment on the random slope structure above, we have altered the supplement such that it splits the output for every lmer simulation in non-singular and singular fits and shows BFs for both. We only used the non-singular fits as the numbers we rely upon for the sample size determination and are thus reported in Appendix A. For all models we plan to use, the number of singular fits decreases with an increasing sample size.

- *In your Rmd file on page 23 you say that you set the value of condition to 0- I assume this is a typo and it's the interaction that you set to 0, however I can't check because the pdf is rendered such that it runs off the page (it needs to be reknit)*

  It was indeed the effect for interaction that was set to 0. We have now knitted the document as a HTML file again, because in the PDF format it kept falling off the page. Thank you for noticing this.

Group analysis of behavioural outcome measures

- *Reader might not know the abbreviation CLMM – perhaps I missed it if not spell out cumulative link model models is*

  Thank you for noticing this. We have now written out "Cumulative Link Mixed Model (CLMM)" on page 19.

- *BAIN: again, spell out acronym (Bayes informative hypotheses evaluation). More substantially, the reader needs more detail to understand how the process works. The requirements of PCRI RR state:*

  "For studies involving analyses with Bayes factors, the predictions of the theory must be specified so that a Bayes factor can be calculated. Authors should indicate what distribution will be used to represent the predictions of the theory and how its parameters will be specified. For example, will you use a uniform distribution up to some specified maximum, or a normal/half-normal distribution to represent a likely effect size, or a JZS/Cauchy distribution with a specified scaling constant?"

  You need to do this for each BF analysis in the paper – it isn't sufficient just to refer us to other papers for details. I am not very familiar with the Bain approach (and do please correct me if I have misunderstood) but my understanding is that for an H1 like beta1>0 the model that represents the theory is a normal (?) distribution which has a parameter for the variance which is set using a fraction

of the data. Can you make these details clear including explaining how process by which the parameters are set based on the data.

Thank you again for noticing, we have now written out the acronym for Bain as well (page 19). We have consulted with prof. dr. Herbert Hoijtink on this point. To make the details of the process by which the parameters are set clearer, we now added information and equations to the manuscript (section 2.5.1. on pages 18 and 19) stating how the prior and posterior distributions, and the BFs are calculated in Bain. This same information also holds for the use of Bain in the mediation analysis. We have placed a reference to section 2.5.1. there as well.

- *Please give details of the sensitivity analysis.*

The sensitivity analysis in Bain is always performed by varying the size of the fraction *b* of information in the data used to specify the prior variance from 1 × *b* (default), to 2 × *b*, as well as 3 × *b.* If the BF does not substantially change as a result of this, we can conclude that the results are robust (Hoijtink et al., 2019, pp. 548-549). We have now added this clarification to the manuscript where we discuss analyses with Bain, as well as in Appendix A.

- *Sample size analysis: As noted above, if this is considered a key hypothesis in the paper, you need to show that your planned sample has a reasonable chance of finding evidence for H0 if H0 is true, and for H1 if H1 is true. Is this a place where you could use the pilot data? That is, extract values from that data and generate random data from those so you can see % of samples with various N where you get evidence for H1, and then simulate data using the same parameters except changing the key parameter to 0 and seeing % of samples where you get evidence for H0?*

For the behavioral data, we have not included this as one of the key analyses (in green in Appendix A), thus we have not done a sample size simulation analysis but instead chose to rely on the results coming from earlier studies and our pilots. We did work on simulations for this but generating ordinal data for the rating task caused too much programming difficulty. We did run simulations for the TDT, which we chose to not include in the end because it is not a key analysis. These simulations did show a robust effect, for both H1 and H0. Please reach out if you would like to see these simulations to verify this. However, we do not wish to include it in the manuscript because it is not a key analysis for the sample size calculation.

Correlations between neural and behavioural data

- *You say you use the default prior in JASP with k=1, but as per the requirements I state above, you need more explicitly spell out what distribution is used and what its parameters are. Your use of this distribution needs to be justified and, as I pointed out above, a concern in using default priors is that it is biasing you to find evidence for the null. I would suggest that values from the pilot or previous studies where you found evidence for these correlations[1] would be your best bet in terms of getting ball-park estimates of the size of correlation you expect. You can also report robustness region, or some other sensitivity analysis.*

Thank you for pointing this out. Our initial understanding was that the default prior provided equal probability for all effect sizes and was therefore the most conservative to use. However, after consulting with our statistics advisors, we concluded that it is reasonable to use informed priors in this case as they do indeed increase the chance of finding evidence for H1 if an effect exists. Again, the behavioral tasks and also their correlations with the neural data are not in our critical analyses for sample size determination, but this point applies to all correlation analyses in our study.

In general, the prior for correlations in JASP is described by a stretched beta-distribution centered around zero and with a width parameter (κ) of 1 as the default. For instance, a prior weight of 0.5 generates a beta(2,2) stretched from -1 to 1 (because 2 = 1/0.5; the width is inversely related to

the parameters of the beta). The width parameter for the beta distribution in JASP can range between 0 and 2. In our case, the beta distribution is cut in half at 0, because we only hypothesize positive correlations. We noticed this was not yet included in the manuscript and have now added this information. A prior width of 0.5 is reasonable for medium effect sizes, and a beta prior width of 0.75 suits large correlations of .6 or greater. To further illustrate this, we have added an image to the manuscript (Figure 4, page 22) with these distributions plotted.

For the neural-behavioral correlations, which now fall in the range of medium effect sizes (0.32 for rating task, 0.42 for TDT), we have chosen the prior $\kappa = 0.5$ and re-ran the BFDA simulations using that prior. We have also altered the manuscript to include this information.

Sample size analyse:

- *Note: If I have understood your code correctly, I think that the test of whether we can find evidence for H0 if it is true is covered in section labelled H0 which is on page 25 and page 26, and the tests for when H1 is true are on page 27 - my comments relate to those*

  When looking for evidence for H1/H0 it isn't clear to me why you have focused on an analysis with a maximum of N=150, when the N you are planning to use is N=105? (Maybe the justification of sample size is based on the ASN, but if ASN=x that just tells us that on average a boundary would be crossed (and thus data collection stopped) when N=x, but it doesn't tell us what % of runs we would correctly find evidence for H1/H0 if data collection always stopped at N=x. (Again, it's great to have explored the implications of using different sample sizes but, as above, what we need to know to assess the planned study is the % of runs in the simulation where - using the procedure you are going to use with a particular maximum - you find evidence for H1 when it is true and H0 when H0 is true).
  Note that if you change to using informed models of H1 for each hypothesis, you will need separate analyses testing for evidence for H0 when H0 is true for each hypothesis.

  Thank you again for this comment. As we pointed out in response to your similar point above, we initially performed all our simulations on sample sizes until a max of 150. From there we looked at what was feasible and what the updating sample would be. We have now performed all BFDA simulations on 45-105, with an updating sample of 15 participants. We also ran H0 correlations with the same informed priors as the H1s. This is exactly the same as what we will do in the experiment. We have altered this in Appendix A as well.

Analyses of individual differences in statistical learning: Correlations

- *You do some simulations looking for the possibility of detecting small, medium and large effects. However, the pattern of findings results from the fact that you are using a default model of H1, which as noted above, is likely to be testing the theory of a large effect. So you are simulating data where the effect is (e.g.) small, and then computing a Bayes Factor which tells you if this data this more likely under a theory where there is effect is large, or one where there is no effect. It isn't surprising that the answer to question is often that the null is more likely. This demonstrates why it is important to use a model of H1 that roughly matches your expectation of effect size in this scenario. Again, perhaps you can use values from the pilot/previous studies to get a ball-park estimate? And then this can be used both as a parameter in the model of H1 and in the simulations where you generate data under the assumption that H1 is true.*

  Under your comment on the neural-behavioral correlations, we explained what the prior distribution is for the JASP correlations. We still performed the BFDA simulations on the small, medium, and large effect sizes, but now with suitable priors for each. We then used the pilot data as an indication of what we can expect. However, we must note that sensitivity analyses show that most of this pilot data is not robust to different priors (see the JASP supplement with the sensitivity analyses for the student pilot data of individual differences). We have now performed our sample size BFDA simulations with these adjusted informed priors and included these results in Appendix A. Our

simulations still show that small effect sizes are not feasible to detect in our project. However, such small effect sizes are also theoretically not interesting, so that is not what we are aiming for. We have included this argumentation as well in Appendix A. We do have reasonable chances to find evidence for medium and large effect sizes. So, for the rhythm tests (SSS, CA-BAT, and PROMS) we expect large effect sizes as they should essentially measure the same concept, and we will use $\kappa = 0.75$. With regard to the other correlations that are not part of our critical analyses, we will use $\kappa = 0.5$ because that places less weight on large effect sizes and has relatively more weight around 0.

Analyses of individual differences in statistical learning: Mediation analysis

- *You say "We will, however, only add tasks as mediators that are significantly correlated with the SSS task in the correlation analysis between all tasks above" but under the Bayesian approach you are taking you aren't testing for significance. Do you mean that if BF >6? What if the BF doesn't meet this criteria but also doesn't meet criteria for accepting the null?*

  We will only run the mediation analysis if we find correlations between tasks and the SSS task that are *BF* > 6. We have now included this more clearly in section 2.6, stating that we will otherwise separately correlate these tasks with the WLI instead. However, the correlations of the rhythm tasks with the SSS task are part of our critical analyses. So, if we do not find conclusive evidence there, we will do another sample update. If we reached our maximum sample size and still have no evidence for correlations between rhythm tasks and the SSS task, we will separately correlate the tasks with the WLI (see pages 24-25 in the manuscript) and only test for the direct effect of the SSS task on the WLI without mediation. However, in our pilot we already found large effect sizes for the SSS task and rhythm tasks, as well as the rhythm tasks among each other (see the JASP correlations supplement). This is not surprising, as especially the PROMS and CA-BAT aim to measure the same concept "rhythmic ability", and the SSS task entails behavioral rhythmic entrainment (albeit more speech-focused).

- *Computational of Bayes Factor and sample size analyses: I got quite confused here. In the paper you say you will use Bain. As above, in that case you need to have details about how H1 is computed under that approach. But what is the relationship between using Bain and the analyses in the script? In the script, you talk about using Dienes heuristic to test the theory that there is a direct effect using the heuristic of a uniform distribution from [0, totaleffect] (which seem quite different from the Bain approach?), and then further confusingly, you actually show a simulation where you use a uniform with a distribution of [0,totaleffect=1] (whereas– as I understand it- Dienes's heuristic was suggesting that totaleffect should be a value that can be extracted from the dataset). Finally, I am confused by why the simulations involve looking at possibility of finding evidence that SSS predicts WLI in a simple regression, not looking at the situation where there is mediation? Is there a justification for this being reasonably equivalent in terms of power?*

  Our apologies for the confusion on this point. We used the total-effect heuristic mainly for its argument stating that the total effect is equally large as the mediated effect. This is the case because if there is mediation, it takes away from the direct effect. Thus, in terms of power, planning for the direct effect is reasonable. However, as you point out, we have made a mistake in how we used this heuristic with the informed prior by using the maximum value of 1. We were unsure before on how to implement Bain in the simulation code, which is why we used the informed H1 based on our pilot results for this power simulation. However, for the analysis of the real data we would indeed like to use Bain. We have now managed to implement Bain in the simulation code (see the simulation supplement), which shows that we have reasonable chances to find evidence for both H1 and H0 (see Appendix A and the simulation supplement).

Footnote:

[1] If you want to use pearson r value from a previous study to inform your model of H1, I am not sure how you do it in Jasp, but FYI you can use Dienes calculator as you did for the other analyses. You will have (I) your

predicted r value (2) your r value from the data (3) your df of the data. You use Fisher's z transform on (1) and (2). Your model of H1 can then be a half normal with sd set to (1) and the model of the data is a has mean (2) and SE = 1/squareroot(df - 1) (see http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm) I am not sure if/ how that works with the BDFA.sim() function you found though.

Thank you for this footnote comment. However, the BFDA function requires us to insert a predicted *r* value and a kappa prior as in JASP. This is on the one hand convenient, because it is the same as what we will use for the analyses, but it required us to find out more about the beta distribution used in JASP. See the comments above and the information added to the manuscript on this matter for clarifications.

Reviewer 1

Thank you for making the changes to the manuscript. I found the manuscript a lot clearer and had a much easier time following the different predictions. I am also glad you decided to use a continuous analysis approach for the different types of synchronizers.

I am not an expert in Bayesian analyses, but from my understanding, the sample sizes are now well-justified, and I would be happy if the study goes forward as planned.

Very minor comments:

- *Figure 3 seems that it was made by taking a screenshot – as a result the nonword is underlined in red (as Microsoft Word does for "misspelled" words).*

- *For Figure 4, would it be possible to add a note that specifies again which test measures what (theoretical) concept? As a reader (who is not very familiar with how for example rhythm is typically assessed), it is hard to keep track of all the different abbreviations.*

Thank you for your positive assessment of our revised manuscript. We have now implemented the requested changes to the figures. We remade figures 1, 3, and 4 with Canva (https://www.canva.com/). This software is easy to use and produces nicer figures than PowerPoint, which we previously used for these images. We also aimed to make figure 4 more conceptual, denoting the mediators as "rhythmic ability" and specifying the tasks measuring this below the concept.

Reviewer 2

- *The authors have addressed all my concerns in the revision and modified their plans appropriately, and I am therefore have no further concerns and am happy to see this accepted. I also think the switch to incremental data collection using the Bayesian analysis technique is very nice*

  Thank you very much for your positive assessment of our revised manuscript. We are also very happy to read that you like the plan for incremental data collection.

We would like to thank you again for your very helpful comments on the previous version of the manuscript, and we hope you will be able to accept the revised version with the responses to your comments and questions provided above.

With kind regards, also on behalf of Marijn Struiksma, Laura Batterink, and Frank Wijnen,

Iris van der Wulp