University of Oslo
Department of Linguistics and Scandinavian Studies
Dr. Timo Roettger

Date: January 31, 2023

Dear Dr. Sbarra,

This letter is to accompany our final submission *Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human speech analyses*.

With our submission, please find (a) our manuscript (RR.pdf), and (b) a detailed response letter, addressing all comments and suggestions by the two reviewers (find below for convenience).

In the hope that our submission will find your approval,
sincerely yours,

Dr. Stefano Coretta
Dr. Joseph Casillas
Dr. Timo B. Roettger (corresponding author)

Note: manuscript pages as mentioned here refer to the pages of the manuscript PDF and not to pages of the proof.

Comments by Julia Strand:

• I'd encourage you to use more descriptive names/include a more detailed readme in the data on OSF. I had trouble making sense of how the data were stored/organized. I see more notes in the manuscript itself, but it would be great to have just a bit more on OSF. That might also remove some of the need to have so many links in the paper.

We have updated the README files contained in each directory of the github repo, which corresponds with the "meta-analysis" component on OSF. Additionally, we have also included a brief description in the wiki section of every core component on OSF. Specifically, we describe the contents of the components (i.e., directories and files) and their relevance to the MSA project.

• Bottom of p 24: The paper does a great job of explaining why asking to report standardized effect size isn't a good approach, but I don't entirely understand what you did instead. Is the paragraph that begins "The coefficients of the critical predictors" explaining that? Or is that another process? Just a bit more detail would be helpful for readers (like me!) who don't have experience with this procedure.

In the revised manuscript we explain the standardisation procedure in the section "*Meta-analytic estimation*". We have included the relevant text below for your convenience:

> "Since the variables used by the analysis teams might have substantially differed in their measurement scales (e.g, Hertz for frequency vs. milliseconds for duration) which was indeed the case, we have standardized all reported effects by refitting each REPORTED MODEL with centered and scaled continuous variables (*z*-scores, i.e. the observed values subtracted from the mean divided by the standard deviation) and sum-coded factor variables."

• This may be a question for the editor, but I wonder about the level of detail in some sections (looking at Model specification section). Given that all the modeling information is available in the open data/code, I wonder whether everything in the paragraph that starts on line 18 of page 28 is necessary.

We are impartial to either option (keeping the information or moving it somewhere else), so we leave this decision to the discretion of the editor. Given that we want to share divergences from the original preregistered plans within the manuscript, we think giving the reader access to a detailed description of the analysis is helpful.

• p29, lines 35+: Great that you ended up with so many teams that met your criteria! Can you say more about whether the exclusion criteria were part of the stage 1 submission or whether those decisions were made after the fact? I’m also wondering whether there was anything systematic about the teams who dropped out (especially given the content in the first full paragraph of p. 43).

We now explicitly state that some of the exclusion criteria for teams were not preregistered:

> “Note that due to the unforeseen variability across teams, the latter exclusion criteria were not preregistered and were applied after having seen all analytic strategies. ”

We did not notice anything systematic about the excluded teams beyond their choices of analyses.

• A gentle suggestion: I wonder whether there is a way to visualize the variability in the approaches that different teams took. Table 1 shows breakdowns within a variable (e.g., outcome type) but not breakdowns across variables. For instance, is it the case that teams that tended to use f0 as the outcome also tended to use the segment as the temporal window? Something like a Sankey diagram might be helpful to show the extent of the variability.

Great suggestion. We have created a plot that shows the connection between choices related to outcome variable, measurement window and operationalization. We have added it to the supplementary material.

• Figure 4 is great—really helpful for showing the results. I wonder whether there is a way to distinguish the rows in plot C as being about the researchers (e.g., years after PhD), what is included in the models (e.g., duration), etc. I think I’d rather see the factors grouped by category (e.g., everything about the researchers themselves first) rather than sorted by order.

We now group the variables in the forest plot into predictors related to (1) temporal window, (2) outcome variable, and (3) team/analysis.

• In the “lessons for the methodological reform movement,” it would be great to see more about preregistration. The content on p. 43 (earlier) is highly relevant to researcher degrees of freedom, but it would be great to see more in the lessons section. This seems like an excellent demonstration of why preregistration is useful: there are many reasonable approaches researchers could (and do) take, and they may lead to different outcomes.

We have added a sentence about the usefulness of preregistration in certain contexts, but do consider its value limited in light of the fact that the theoretical landscape does not allow for precise hypothesis formulation.

> “Preregistration, i.e. a time-stamped document in which researchers specify how they plan to collect their data and/or how they plan to conduct their confirmatory analysis, is can be a useful tool to safeguard researchers against the urge to explore many

different analytical paths before choosing the one that, in hindsight, seems most justified.
However, as long as the theoretical landscape does not allow for more precise hypotheses, the value of preregistration is limited and we need to find ways to appropriately calibrate the confidence in our claims:”

Comments by Matt Goldrick:

The changes to the refitting procedure should be summarized in the manuscript text (and in the change log for completeness).

This has now been added to the ms and changelog.

The only divergences I noted that were not specified either in the manuscript or these Google Docs were treatment of divergent transitions in the Hamilton Monte-Carlo sampling procedure. This should be specified in the text as well. I do not think this will alter the results in a significant way.

The following sentence has been added:
“All models converged ($\hat{R}$ was approximately 1) and only some had a small number of divergent transitions which were not deemed problematic (range 1-156).”

Page 28: last sentence of first paragraph "this we decided against"-> thus we decided against

Fixed.

Page 29: I requested and received access to the comment history for the manuscript (footnote ||). Prior to publication this should either be made open to any viewer, or the footnote should clarify that users must request access. While it did not alter my interpretation of the findings, this provided an interesting picture into the collaboration! It was great to see how various points raised by co-authors influenced the writing of the paper.

We have now made all linked documents stored on Gdrive public.

Page 31: Second paragraph under "statistical analysis" -- r code shown, rather than upper value of range

Fixed.

Page 46: Third full paragraph "are relateD to mathematically formalizing verbal…"

Fixed.

Page 46: It might be good to note that mathematical/computational modeling, while important, currently works in spaces that are much lower in dimensionality than the system we are measuring. We are still stuck with the issue of how to relate an abstract modeling space to the measurement space.

We added a footnote to clarify this point:

> "Although conceptually promising, in their current state, such formalized models typically work in spaces that are much lower in dimensionality than the complex systems in which we measure. Thus, future research should spend resources on attempting to quantitatively relate the abstract theoretical space to the complex measurement space."

Page 50: CReDIT URL should be: https://github.com/many-speech-analyses/many_analyses/blob/main/figs/credit-taxonomy-all.png

Fixed.

Divergences between RR and Manuscript:
Re-fit each analysis
Plan: use Bayesian regression using centered, scaled continuous variables and sum-coded factors
The text (p.25) provides a link to the analysis workflow, which differs in a few details:

- Typicality, if categorical, is dummy coded (others are sum-coded).
- "If analysts used model selection, use the model which analyst eventually chose unless final model did not include typicality. In that case, chose simplest model with typicality and their chosen critical predictors"

We have specified the changes to the protocol:
"Relative to the registered protocol, we made minor changes to the refitting procedure, specifically file and variable naming conventions and the use of treatment contrasts instead of sum coding."

We would like to thank both Julia Strand and Matt Goldrick for their insightful comments.