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

Dr. David A Sbarra
Advances in Methods and Practices in Psychological Sciences

Re: “Revision of Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human speech analyses”

Date: November 7, 2022

Dear Dr. Sbarra,

This letter is to accompany our stage-2 submission *Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human speech analyses*. We collected analyses from 46 different analyses teams who submitted over 100 analyses.

With our submission, please find (a) a “clean” manuscript (Draft_RR.pdf), (b) a version of the manuscript that highlights each individual change between the accepted Registered Report and the final manuscript (Draft_RR_diff.pdf), and (c) a change-log that details what we changed when, and why (https://bit.ly/3t6wNYd). All materials, including raw data, derived data and scripts alongside the approved Registered Report can be found on a publicly available repository here: https://osf.io/3bmcp/.

After the collaborative editing process, several elements of the introduction and method section have been edited to increase clarity and readability (see (b)). These are all changes in the interest of the reader and they create, what we consider, a much improved reading flow. Please let us know if you or the reviewers think any of these changes alter the preregistered rationale.

Moreover, we encountered two unforeseen challenges to the preregistered analysis plan:

1. While we expected all teams to submit only one model, many teams ignored this part of the instruction (reflecting the general tendency in the field to test one hypothesis with many different models). Models ran by the same team result in dependencies between data points and need to be accounted for with a different group-level structure than originally preregistered.
2. We had to exclude two preregistered predictors:
    1. The predictor “number of post hoc analyses” could not be included because teams did not report any post hoc analyses.

E-mail: timo.roettger@iln.uio.no
www.uio.no

2. The predictor “number of random effects” (i.e. model conservativeness) could not be included because this predictor turned out to be highly correlated with the estimated standard error of the measurement error model (more random effects = higher standard error) and thus distorted the model estimates.

We report all these changes within the manuscript and link to our statistical code to run the preregistered analyses. We also partly report on the differences between preregistered and non-preregistered analyses in the manuscript as we saw fit.

In the hope that this stage-2 submission will find your approval,
sincerely yours,

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

LEGEND

bold and black = author response
light and grey = reviewer comment

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

Response to Dr. Strand's comments:

The reviewers and I agree that this paper could make a useful contribution to the literature and is appropriate for a Registered Report. The reviewers also made helpful suggestions for improvement which you can see below. I'd encourage you to incorporate them in your revision, and also have several suggestions of my own:

1. Reviewer 1 wrote that they would "encourage you however to find a way to go beyond replicating other multi-analysts paper" (comment 10). I think there is a good argument for how you are doing that with this research (i.e., that the input researchers are starting with in your study is even less well defined than numbers in a database), but that argument could be made more strongly. What additional benefit does starting with something as abstract as speech recordings provide?

Thank you for this suggestion. We now explicitly define our goals as two-fold in section 1 (page 3 of the paper, second paragraph):

> "Our goals are twofold: (i) our study conceptually replicates previous many-analyses projects, by probing the effects of different statistical analyses and by assessing the generalizability of published findings to other disciplines (here, the speech sciences);
> (ii) our study extends the scope of inquiry to include flexibility in the operationalisation of complex human behavior (here, human speech).
> This is an important addition in that the increased number of "forking paths" in the "garden of analytic choices", deriving from the many decisions involved in quantification, might reveal a higher degree of variability across analysts than what previously observed, thus giving us a more realistic estimate of variability across analysts."

We now make explicit that our project might provide a more realistic upper bound of how flexible data analysis can be given degrees of freedom in both measurement and statistical analysis (page 14, last paragraph):

> "The complexity of speech data constitutes the ideal testing ground to assess the upper bound of analytic flexibility that social science might face across disciplines."

We further reworked the manuscript by emphasizing these goals throughout the introduction.

2. This new paper is likely to be relevant: https://journals.sagepub.com/doi/pdf/10.1177/23780231211024421

Thanks for pointing out this paper. We have now integrated the conceptual take-away of the article in our discussion.

3. On page 15, when describing the degree of uncertainty around effect variability as driven by number of teams, you mention that 10 teams should be sufficient, but then also mention 200 teams indicating interest, and that even 50% of that would be sufficient. It would be helpful to be a bit more explicit in justifying the minimum number of teams that you would consider appropriate for these analyses.

In order to achieve our main goal, i.e. estimating variability among teams, we consider a minimum sample size of 10 teams as sufficient (as estimated from simulations). Given the exploratory nature of our study (in absence of a clear precision target), however, we will sample as many analysts as possible. We have explicitly added this reasoning to the manuscript now. (See *Phase 1: Recruitment of analysts and initial survey* section, page 16, second paragraph.)

4. If I understand correctly, the speech files to be analyzed are not from a previously published paper. I assume this was done to avoid the possibility that a team could try to replicate those analytic methods and simply reproduce the original analyses? It would be helpful to be more explicit about whether the methods you employ are typical in the discipline.

Results of this research project were neither published nor publicly presented and are stored on a private OSF repository, thus potential analysts will have no access to replicate the original analysis. We added this information to the text as a footnote (denoted with "*" at the bottom of approximately page 11 of the revised manuscript). The dataset and methods employed in the unpublished study are quite typical for the field. We have added a mention to this in the text.

We would like to thank the editor for considering our proposal and those insightful comments

Responses to Reviewer 1

The presented research examines to what extent researchers and research teams differ in the operationalization of primary data. The research setting is in the area of communication and speech analysis. I am not familiar with the literature and conventions in this research area. I have tried to formulate my comments to the best of my knowledge, but they come from an outsider of your field. I appreciate the presented data and links to the questionnaires as well as the informative website and video, which will be used for recruitment.

> We are glad the reviewer found the materials useful and appreciate their outside perspective on our study.

1. The main research question appears to be whether researchers also in this setting would differ in their operationalization of variables. The greater the number of analysts involved, the larger the diversity of their backgrounds, trainings and expertise, thus the larger variation in operationalization would be expected. That is also the less guidance is given by the research team, the more researchers will rely on their own experience, expertise, training and priorities to assess the speech samples. It appears to me that the main justification of the research is in '...assessing the variability in the decisions regarding the quantification of the measured behavior'. I would encourage you to sharpen this research rationale, and explain why such an assessment is indeed worthy the massive investment of a crowd effort.

> Previous multi-analyst studies repeatedly failed to provide any evidence that the "backgrounds, trainings and expertise" of analysts explains observed variability, thus it remains an empirical question whether individual characteristics of analysts really affect the outcome of their analysis.
>
> We have reworked the introduction to emphasize our goals (page 3 of the paper, second paragraph):
>
> > "Our goals are twofold: First, our study conceptually replicates previous many-analyses projects, by probing the effects of different statistical analyses and by assessing the generalizability of published findings to other disciplines (here, the speech sciences). Second, our study extends the scope of inquiry to include flexibility in the operationalisation of complex human behavior (here, human speech). This is an important addition in that the increased number of "forking paths" in the "garden of analytic choices", deriving from the many decisions involved in quantification, might reveal a higher degree of variability across analysts than what previously observed."

2. The dataset is in German, the research teams are recruited internationally, using an informative website with a clear instructional video. I'm wondering to what extent outcomes are affected by the extent to which teams or individuals participating are familiar with the German language. How are you planning to code

team characteristics such as familiarity with the German language (items that you included in your intake questionnaire) if multiple people are involved in team, with varied experience.

> We acknowledge the concern that familiarity with the target language might affect results. We actually think that using a data set that is *not* based on English speakers reduces the average familiarity of analysts with the specific phenomena under investigation and thus allows for a less biased assessment of the data. In other words, we see this as an advantage rather than a disadvantage. The language sciences are still very anglocentric, leading to a high familiarity with patterns in English. A dataset on English speakers might amplify analytical decisions based on prior beliefs. The question of familiarity with cultural or linguistic properties of a dataset is in itself an interesting research question but outside the scope of our present study. From a technical point of view, speech scientists are trained to analyse acoustic data of unfamiliar languages, so we do not anticipate people having difficulties analysing a German data set. Moreover, we believe that analysts who are "blind" to the meaning of the recordings are arguably less biased. In order to guide analysts, our metadata provide detailed descriptions of the relevant components of the data set. We also added a couple of items to our surveys that assess this information: We are now asking analysts to specify their proficiency in German and whether they have worked on German speech data before. (See question 13 of the intake form, available here: https://osf.io/w7db5/).

3. In the intake form I saw that you ask about statistical experience, and researchers' current level, but I am wondering if asking for individuals' discipline, their prior experience with speech analysis, with software such as Praat, perhaps also their level of other languages (linguistic competence as a whole) may be relevant.

> Thank you for this suggestion, we have added an item assessing analysts' familiarity with Praat (which is the discipline standard across speech scientists): https://osf.io/w7db5/

4. You could think about separating a team registration form (where also after initial registration researchers could be added to a team) - and separate this from an individual sign up form, where individuals' expertise is captured.

> This is indeed how the forms are structured. The intake form must be filled out by every team member individually.

5. A clear idea on how you will aggregate measures for team level analysis would be helpful for later analysis.

> We have expanded the text explaining the procedure (page 22, first paragraph):
>
> > "To obtain an aggregated research experience score and initial belief score for each team based on the members' individual scores, we will calculate the mean

and standard deviation of these predictors for each team. These will be entered in the model formula as a measurement-error term (`me(mean, sd)` in brms). The expedient of using a measurement-error term (which includes the teams' standard deviation) ensures information about within-team variance is not lost (which would be the case if including the mean only)."

6. It may be helpful for you to have a post-hoc questionnaire identifying who in the team did the particular analysis, and to get an idea of if and how the team worked together - or whether work was strictly split.

Both of these points are great suggestions. There are many different aspects of the analysts and the way they form teams that might have an influence on the results in subtle (or not so subtle ways). However, we would have little to say about how these dynamics might possibly explain variation in an a priori manner. Moreover, given the fact that previous multi-analyst studies repeatedly failed to find that individual characteristics of analysts explain variation (which might be due to insufficient sample sizes), we fear that these complex measures might not yield informative results.

7. On page 20 of your manuscript it sounds like you are planning to recruit peer-reviewers separately, hence people who have not participated in the initial analysis, and, should this number not be sufficient, recruit reviewers from the analysis teams.
The experience from our project in the past showed that peer reviewers benefitted from having worked with the data themselves. You may think about working either with participants as reviewers - or working with and briefing outside reviewers. Given a sufficient number of analysts you could also think about randomly assorting participants into analysts or reviewers.

This is a great comment. We discussed this issue and decided that we will let analyst teams randomly review other analyses after they have submitted their analysis. This way we can ensure that every reviewer has a similar level of familiarity with the data. We have amended the manuscript accordingly.

8. It appeared to me that the initial focus and motivation for your manuscript was to test for differences in operationalization of primary speech data, yet the stages of your analytical approach then focus more on the statistical operationalization. Please excuse my naïvety of common analysis in your research area. I am wondering if what you are trying to do with your research isn't so much whether researchers differ in their approach to operationalization (which will likely be the case, particularly with increased number of participants). Could it be that you are trying to find out which statistical approach to linguistic analysis would be capable of distinguishing acoustically modified utterances to signal atypical word combination!?

And you are using a crowd-based approach to assess whether such a technique can be found, which would have implications for the automated assessment of speech. So researchers would be tasked with the question 'which acoustic measures can be used to identify modifications in utterances for atypical word combinations?'. If such an approach cannot be found by any team, then your alternative hypothesis may be true, that speakers do not (or at least not in a uniform way) modify utterances for atypical word combinations. The team and expertise variables that you are collecting could then provide input for

interesting analysis, identifying which team compositions and expertise was necessary to identify such an approach. You could also think about holding back some of the data so that the found approach could then be tested with a new dataset, similar to how computer scientists separate data used to training models from data used to test the established model.

> Thanks for sharing these thoughts with us. The project is not meant to focus on the research question per se, in the sense that we are not really interested in knowing the answer to the research question. The research question is representative of the research conducted within psycholinguistics and speech sciences and functions only as a test case here. We fully agree with Auspurg & Brüderl (2021) that crowd-sourcing analyses to estimate parameters of estimates might still be premature before we have fully understood the dynamics that lead to analytical flexibility.
> Our study attempts to contribute to such a better understanding and focuses on estimating the analytic flexibility both at the level of operationalisation *and* statistical analysis (rather than just at the level of statistical analysis). We clarified our goals in several parts of the manuscript (see above).

9. As you suggest using meta-analytical techniques for the analysis of the different studies, please have a look at this post (http://datacolada.org/73) describing why this may not always be a valid choice.

> Our meta-analysis will not be used to establish a best estimate of a true effect, but rather to quantify variability across individual analyses in relation to the crowd's overall estimate. We think the concern articulated in the (very informative) article from Datacolada on study-internal meta analyses does not apply to our context here unless we misunderstand the article or the reviewer.

10. In short, I believe that your research can make an important contribution, and is of particular relevance to your field. I would encourage you however to find a way to go beyond replicating other multi-analysts papers. Perhaps the approach mentioned in point 8 can give some food for thought about exploring this research through a different lens, while retaining the overarching research setting and research question. Regardless if you follow that approach or another approach, given that this is a large-scale study, the unique contribution of this particular research effort, and the justification for using a crowd-sourced approach would need to be stronger to justify the massive investment of research time dedicated to this project.

> We have clarified that the goal of the project is to investigate the degree of analytic variability as derived from flexibility in both statistical analysis and operationalisation. We believe this to be the main point that sets this project apart in respect to other many analyst projects.

We would like to thank the reviewer for their insightful comments and suggestions.

Response to Reviewer 2

For the meta-analytic estimation procedure, should the authors specify a random seed (which may impact the replicability of their MCMC calculations etc.)? The code suggests there is one ("my_seed") but it's not declared anywhere in the code.

> Thank you for the suggestion. We have specified a random seed both globally (with set.seed()) and in individual models (with the seed argument of brm()).

Question about the task: The authors focus on the replicability of analyses and how this will be addressed by a large, crowd-sourced analysis approach. In earlier framings of the "replicability crisis" a lot of concern was about p values and invalid statistical inferences. Does the dataset here represent a lower bound? That is, when participants analyze the speech dataset that is provided, will they be approaching the data as they would their own data? Said another way, will any variability in the final results about the multidimensional be explained by researcher characteristics (e.g., some folks take open science very seriously, approach this as a test, etc.)? I see that the variable of the researchers' belief in the general hypothesis of the original study is included as a control variable, which is great foresight.

> If we understand the reviewer's comment correctly, our aim is to obtain an estimate of inter-team variability (in the manuscript, the "Meta-analytic group-level standard deviation: ($\sigma_{\alpha t}$) standard deviation of the group-level effect of team returned by the meta-analytic model") in relation to the estimated effect of typicality; variability which is derived from flexibility in both statistical analysis and operationalisation. While our main goal is the estimation of such variability, we will also conduct an exploratory analysis of the influence (or lack thereof) of a set of factors related to analytic choices and researchers' characteristics on the variance (in other words, we will check whether these factors are correlated to the inter-team variance and to what extent.)

Since there is peer review between contributors (researcher groups), I am wondering what value there may be in what could be considered very poor analyses? Will the authors find their question answered if, "this analysis was conducted reasonably and with sound statistical guidance", especially on the replication crisis front? Perhaps there is something in the meta-analysis on this that I have missed.

> Thanks for pointing this out. We hope that by rephrasing and making the objective of the study more clear (i.e. estimating inter-team variability because of statistical and operationalisation flexibility) we have addressed this point.

Minor comments:

Line 23-26 pdf page 2: Does the sentence, "These researcher degrees…" need to be part of the next paragraph?

> Amended.

Line 13 pdf page p. 10: I would say speech science lies at the “intersection” rather than the “heart” -- as a turn of phrase it is my intuition that “heart” implies that speech science is a core aspect of these interdisciplinary fields. Not that it couldn’t be! But any multidimensional language processing (incl. signed languages) is going to be this interdisciplinary.

Great point. We changed the wording accordingly (page 6, Operationalizing speech, first paragraph):

> “Research on speech lies at the intersection of the cognitive sciences, informing psychological models of language, categorization, and memory, guiding methods for diagnosis and therapy of speech disorders, and facilitating advancement in automatic speech recognition and speech synthesis.”

Line 6 pdf page 11: The authors introduce “speech categories” quite suddenly. I think it would be good to preface how this relates to the “relevant units” introduced in the previous paragraph so that we can anticipate discussion of phonetic segments, which some readers may not understand.

We have reworded the section. Instead of talking about speech categories, a rather vague term, we now refer back to the linguistically relevant units that we introduce above (page 7 end of second paragraph and beginning of third):

> “Utterances stretch over several thousand milliseconds and contain different levels of linguistically relevant units such as phrases, words, syllables, and individual sounds. The researcher is thus confronted with a considerable number of parameters and combinations thereof to choose from. From a phonetic viewpoint, linguistically relevant units are inherently multidimensional and dynamic: they consist of clusters of parameters that are modulated over time.”

Line 20, pdf page 11: Maybe reorder “led” and “let” to match the /t, d/ for the contrast example.

Amended.

Line 20, pdf page 12: I would consider adding citations for these measures just so readers can follow along so those who are less familiar with the definitions of the terms have the chance to look the constructs up.

We have added a citation to the classical “Acoustic phonetics” by Stevens.

Line 41, pdf page 14: “and making” should probably be “and make”

Amended.

Line 9, pdf page 15: What do the authors mean by the “indexical potential of speech” here?

We removed this sentence fragment as we felt it interrupted the flow of the argument.

Line 5, pdf page 17: Realizing that the term is somewhat theoretically burdensome, could the authors please define “focus” here – perhaps in the simplest of terms? This would go a long way toward explaining why an “entire noun phrase” can be in focus.

We have added a definition and exemplified the concept on concrete examples of the experiment (page 13):

> “The two sentence prompts were used to create a focus contrast between the competitor and the target object. Focused units denote the set of all (contextually relevant) alternatives (e.g. Rooth 1992). Concretely, a focus contrast marks one or more elements in a sentence as prominent, by different linguistic means depending on the language. For instance, if the competitor and target objects differ but their color did not (e.g. *yellow banana* vs *yellow tomato*), the noun is said to be in focus (the Noun Focus condition, NF). If the objects are the same but differ in color (e.g. *yellow banana* vs *blue banana*), the color adjective is in focus (the Adjective Focus condition, AF).”

We would like to thank the reviewer for their insightful comments and suggestions.