March 5, 2023

Dear Dr DeBruine,

We are delighted about your enthusiasm regarding our project and thank you very much for your helpful comments.

As recommended, we have simulated data and written our analyses code and realized how useful it was to rethink important aspects of our proposed analyses. We take it as an instructive lesson for good scientific practice that we will certainly apply in future research as well.

Below, we are answering to each of your comments and provide information about the changes in the report itself. Note that we kept track changes to help you localize the proposed changes.

Analysis scripts

I'd like you to include R scripts with your proposed analyses, run on simulated data. You can simulate entirely null effects (just use simulate 45 observations for each perceptual rating for each stimulus with realistic distributions on the Likert scale) or simulate data with a mixed effects structure. I've included some example code below to get you started, based on a faux tutorial.

Please use the simulated data to write explicit analysis code for your planned analyses. It is my experience that describing analyses with prose is too ambiguous for a registered report and leaves all parties open to misunderstanding. For example, I'm not very familiar with MM1 measures and the extent to which they suffer from pseudoreplication; explicit code would be helpful for comparisons with other methods. Please also be explicit about any corrections for multiple comparisons (e.g., include your critical alpha for concluding significance for each test in the code).

If possible, include explicit code and decisions rules for determining if raters should be excluded for inattentiveness (although I acknowledge that sometimes they do things that are obvious only in retrospect, and would absolutely support modification of the criteria is warranted).

We have included our critical alpha for each test in the code and were explicit about corrections for multiple comparisons, as well as about participant data exclusion criteria.

Importantly, we have made slight adjustments to our analysis plan: First, we have explicitly stated that whenever requirements are met, we plan to use parametric methods, aiming at higher power; otherwise, we will use non-parametric alternatives (page 8). We have included these

possibilities in our proposed analyses code. Second, we have reviewed how we plan to treat data from the two testing sessions in our comparison of MM1 agreement between singing styles (Question 1A from Table 1) as well as in our comparison of rankings of singers (Question 2A): for these analyses (as well as for the other proposed inter-rater agreement measures of intraclass correlations and Krippendorff's alpha) we will pool data from both sessions (averaging values of sessions 1 and 2). We have updated Table 1 (page 6) and the text description of our analyses plan (sections 1.2, 1.2.1 and 1.2.2, page 8) to reflect these changes. Third, to evaluate the consistency of MM1 agreement across the two testing sessions (Question 1B), we realized that it would be better to also have a direct comparison of MM1 values in sessions 1 and 2 instead of only correlations between these values, so we are now also proposing a paired t-test (or a Wilcoxon paired test in case parametric tests are not appropriate to the actual collected data; Section 1.2.1, page 8). And fourth, we have specified that we will use a Spearman correlation to compare rankings of singers across sessions (Question 2B - Table 1 and corresponding section 1.2.2).

The simulation method can also help you to do power calculations for analyses that don't have an analytic solution (e.g., in Gpower). I have a tutorial guide to this method using mixed models, but the technique is applicable to any analysis (just write a function that simulates the data, runs the analysis, and returns the p-value of interest; then run this function ~1000 times and report the proportion of p-values less than your critical alpha).

We find the approach of power analyses through data simulation interesting and promising, but were frustrated by difficulties in simulating data with distributions that would be informative to our specific planned analyses. For instance, generating liking ratings that would lead to differences in the amount of MM1 agreement between singing styles; or to different effect sizes in preferences for certain singers (to assess power of Friedman test). This is why we did not change our approach to power analysis in this report.

Rater details

I'd also like a more detailed justification of the raters. It seems probable that consistency of rating things like timbre or resonance would be much more consistent for raters with musical experience (especially choir experience) and some of your results may be entirely dependent on the type of raters you recruit. You state that the subject poolis "mostly lay listeners". It would be good to be more explicit about whether and how raters' experince is expected to affect the ratings and how this might affect the generalisability of your results.

By recruiting participants from the Max Planck Institute's participant database – with varying degrees of music training, but composed mostly of lay listeners – we aim to examine participants with a large range of expertise, which is meant to be representative of a general population. We acknowledge that our convenience sample shares the generalizability limitations of most studies sampling from "WEIRD" populations (White, Educated, Industrialized, Rich, and Democratic – Henrich et al, 2010) and hope for follow up extending our project to other population. As for now, our choice to focus on lay listeners aims at allowing for more generalizability than if we recruited only music experts. However, we agree that expertise might be an important factor in ratings and their consistency, and plan to also collect information about participants' music experience (with the Goldsmiths Music Sophistication Index – Müllensiefen et al., 2014), to be able to explore the relationship between agreement and music sophistication.

Importantly, while it makes sense to expect experts to be more consistent when rating voices, studies suggest that lay listeners are able to evaluate voices in a satisfactory way if suitable scales are made available to them. Using the Geneva Voice Perception Scale (GVPS), Bänziger et al (2014) showed that untrained listeners could evaluate spoken voices in terms of loudness, pitch, intonation, sharpness, articulation, roughness, instability, and speaking rate, and reported Intra-class correlations (ICCs) between 0.22 and 0.81 per feature. Merrill (2022) developed a similar instrument to evaluate singing voices, employing nine different bipolar scales, and collected ratings from speaking- and singing-voice experts as well as lay listeners. She reported ICC measures ranging from negative values for pitch (-.44) and articulation precision (-.25) to higher values for pitch changes (.5), noise (.54) and tension (.57), and a repeated-measures analysis of covariance indicated that expertise did not affect the assessment of singers. Note that low to moderate values of inter-rater agreement were also reported by other studies with expert ratings of voices (e.g., Merrill & Larrouy-Maestri, 2017, where speech therapists rated vocal-articulatory expression of performances of Arnold Schoenberg's speechsong composition "Pierrot lunaire") and music (e.g., Lange & Frieler, 2018, where audio engineers rated perceptual features of contrasting music stimuli). In other words, expertise is not necessarily leading to high agreement. Also note that, at least for the evaluation of pitch accuracy, studies have shown that the effect of expertise is limited, and that lay listeners are consistent when judging pitch accuracy of singing by untrained singers (Larrouy-Maestri et al, 2015) as well as operatic singers (Larrouy-Maestri et al,

2017). We have added a summary of this information to the main text to clarify our choice of sample.

Rating norms

I'd like to recommend you have a look at this paper and see if it might apply to your study:

Taylor, J. E., Rousselet, G. A., Scheepers, C., & Sereno, S. C. (2021, August 3). Rating Norms Should be Calculated from Cumulative Link Mixed Effects Models.https://doi.org/10.3758/s13428-022-01814-7

Thank you for the referral to the Taylor et al (2021) paper. It does raise important points about our handling of rating-scale-generated (ordinal) data as continuous. However, since in the present case we are not focusing on rating norms of per-item aggregated scores of our stimuli per se, we did not change our analyses plans in this regard. In any case, our rankings of singers should produce less biased estimates, as discussed by the authors; and MM1 agreement, despite being based on average ratings across participants, doesn't focus on those mean ratings. We will however keep those concerns in mind, also when discussing our results.

We look forward to hearing from you and to respond to any further questions and comments you may have.

Sincerely,

Camila Bruder, Klaus Frieler, and Pauline Larrouy-Maestri

References:

Bänziger, T., Patel, S., & Scherer, K. R. (2014). The role of perceived voice and speech characteristics in vocal emotion communication. *Journal of Nonverbal Behavior, 38*(1), 31–52. https://doi.org/10.1007/s10919-013-0165-x

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *The Behavioral and brain sciences*, *33*(2-3), 61–135. https://doi.org/10.1017/S0140525X0999152X

Lange, E. B., & Frieler, K. (2018). Challenges and Opportunities of Predicting Musical Emotions with Perceptual and Automatized Features. *Music Perception*, *36*(2), 217–242. https://doi.org/10.1525/mp.2018.36.2.217

Larrouy-Maestri, P., Magis, D., Grabenhorst, M., & Morsomme, D. (2015). Layman versus professional musician: Who makes the better judge? *PLoS ONE, 10*(8), Article e0135394. https://doi.org/10.1371/journal.pone.0135394

Larrouy-Maestri, P., Morsomme, D., Magis, D., & Poeppel, D. (2017). Lay listeners can evaluate the pitch accuracy of operatic voices. Music Perception, 34(4), 489–495. https://doi.org/10.1525/mp.2017.34.4.489

Merrill, J., & Larrouy-Maestri, P. (2017). Vocal Features of Song and Speech: Insights from Schoenberg's Pierrot Lunaire. *Frontiers in psychology*, *8*, 1108. https://doi.org/10.3389/fpsyg.2017.01108

Merrill, J. Auditory perceptual assessment of voices: Examining perceptual ratings as a function of voice experience. *Curr Psychol* (2022). https://doi.org/10.1007/s12144-022-02734-7

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. PloS one, 9(2), e89642. https://doi.org/10.1371/journal.pone.0089642