Frankfurt am Main, 06.02.2024

Dear Dr. Chris Chambers, dear reviewers,

We are very grateful to our reviewers for their valuable feedback. Their input has helped us improve our work, and we believe our revised manuscript now meets the expectations and requirements of a registered report.

We have carefully considered the reviewers' suggestions and have implemented the majority of them. In cases where we have not, we have provided clear justifications for our decisions.

Before addressing the reviewers' comments in detail, we would like to highlight the major changes that have been made to the manuscript in the last round of revision:

- We have restructured the manuscript so that all analyses that were previously listed as “complementary” were moved to an Exploratory Analyses section (2.4.3) of the Methods.

- We have focused our power analysis only on the proposed confirmatory analyses presented in Table 1 (Registered Report Design Planner).

- We removed all mentions of outdated analysis plans for Question 2 from the manuscript.

- We have included a prediction for the expected amount of shared taste for the two speaking styles (Question 1): considering that both adult- and infant-directed speech are “natural” and highly behaviorally relevant, we predict that they will elicit equivalent amounts of shared taste, and propose to test this with equivalence testing. However, we did not feel that we had sufficient theoretical support to make a joint prediction for singing and speaking styles. Even among the three co-authors, we identified different arguments and predictions that could support differential (or not) levels of agreement

between singing and speaking styles. So, even though we would prefer to approach all five styles in an integrative way also for hypothesis 1, we have ultimately refrained from making a joint prediction for the five styles.

- Inspired by a question from reviewer Pat Savage, we have decided to consistently use “mean-minus-one” (MM1) to measure both interrater agreement in Question 1 and interstyle agreement in Question 2 (instead of using Krippendorff’s alpha for Question 2). We computed and compared both agreement measures for our previous data, simulated data, and openly available data from Martinez et al. (2020) and Vessel et al. (2018). We would like to share some of these comparisons with our reviewers:

-

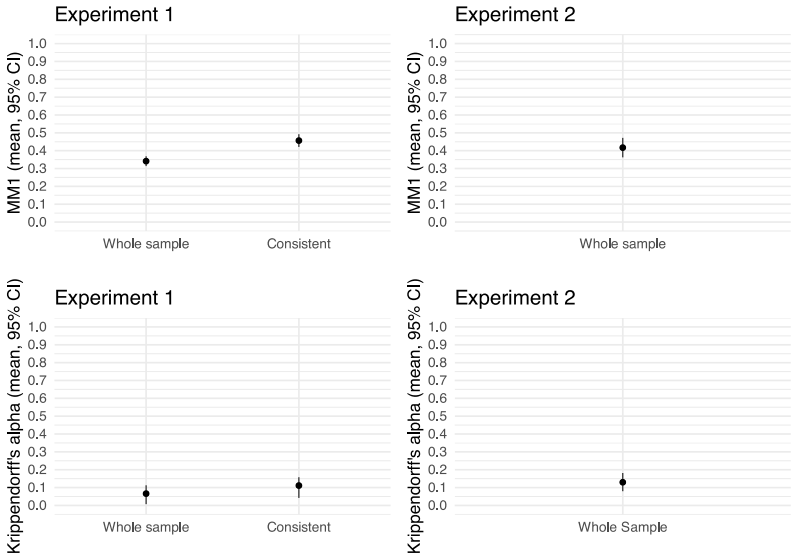

Figure 1: Alternative interrater agreement measures of MM1 and Krippendorff’s alpha computed based on data from Bruder et al. (2023) for liking of pop singing. Error bars represent 95% confidence intervals based on individual MM1 values, or bootstrapped for Krippendorff’s alpha.

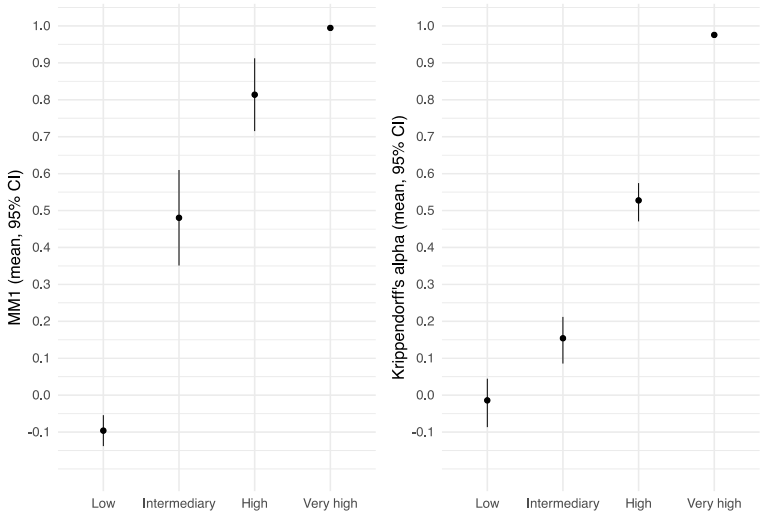

Figure 2: Alternative interrater agreement measures of MM1 and Krippendorff's alpha based on newly simulated data with increasing amounts of interrater agreement. For simplicity, simulations were made for 60 raters rating 22 items, and for only one vocalization style. In the low (null) agreement scenario, all ratings were random. In the intermediary scenario, the ratings given by 30 raters were nearly identical to each other (that is, the same sequence of integer ratings was given for all stimuli, with minimal variability introduced in the data generation process to avoid issues related to perfect correlation), and the ratings given by 30 raters were random. In the high agreement scenario, the ratings given by 50 raters were nearly identical to each other, and the ratings given by 10 raters were random. In the very high (near perfect) agreement scenario, the ratings given by all 60 raters were nearly identical to each other. Error bars represent 95% confidence intervals based on individual MM1 values, or bootstrapped for Krippendorff's alpha.

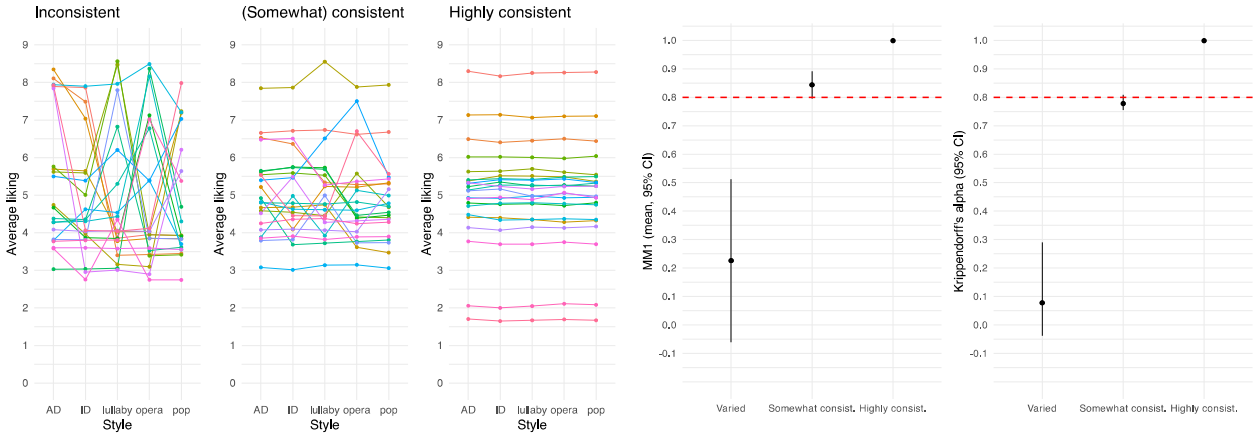

Figure 3: MM1 and Krippendorff's alpha computed as "interstyle agreement" measures, based on simulated data with varied, somewhat consistent, and highly consistent average ratings by singer across the five styles. Error bars depict 95% confidence intervals based for MM1 values, and bootstrapped 95% confidence intervals for Krippendorff's alpha. The dashed red line represents the stipulated threshold of .8 to consider preferences as "highly consistent".

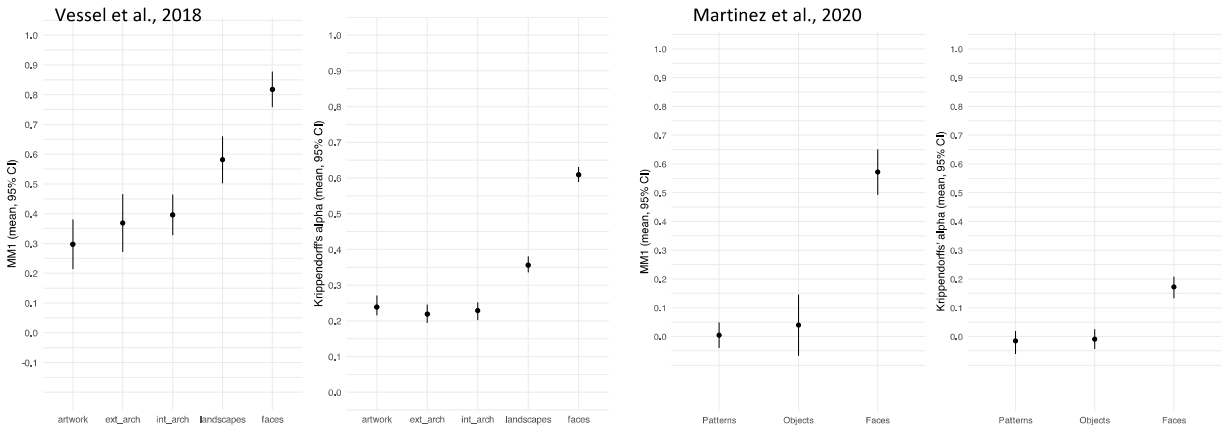

Figure 4: Alternative interrater agreement measures of MM1 and Krippendorff's alpha computed based on openly available data from Vessel et al. (2018) and Martinez et al. (2020). Error bars depict 95% confidence intervals based for MM1 values, and bootstrapped 95% confidence intervals for Krippendorff's alpha.

Crucially, for Question 2 and interstyle agreement of average preferences for some singers, we see that the stipulated threshold of .8 to consider preferences "highly consistent" also seems appropriate when using MM1. That is, MM1 values tend to be higher than Krippendorff's alpha, but they also seem to lead to wider confidence intervals, so most likely, both measures would lead to similar interpretations (and in any case, a finding of interstyle agreement near the stipulated threshold of .8 would need to be properly discussed).

After consideration about peculiarities of each measure (i.e., MM1 can be defined for each participant and measures agreement to the group, whereas Krippendorff's alpha is only defined for groups and measures agreement in a whole group of raters), we ultimately chose to use MM1 for both questions for two main reasons:

- MM1's feature of measuring agreement in relation to the group is particularly appealing for our research on aesthetic experiences.

- Since we relied on theoretical background from the visual domain, it makes sense to also use an agreement measure established in that literature.

We still plan to report Krippendorff's alpha and Intraclass Correlations for both questions, though (as described in Section 2.4.3.2 under "Exploratory analyses"), to contribute to methodological discussions about agreement measures.

Refining our analysis plan prior to data collection has been a long and rich process and a great learning experience, which we certainly wouldn't have had without the registered report format.

Sincerely,

Camila Bruder,

(on behalf of all authors)

References

Martinez, J. E., Funk, F., & Todorov, A. (2020). Quantifying idiosyncratic and shared contributions to judgment. *Behavior Research Methods*, *52*(4), 1428–1444. https://doi.org/10.3758/s13428-019-01323-0

Vessel, E. A., Maurer, N., Denker, A. H., & Starr, G. G. (2018). Stronger shared taste for natural aesthetic domains than for artifacts of human culture. *Cognition*, *179*, 121–131. https://doi.org/10.1016/j.cognition.2018.06.009

Review by Patrick Savage, 14 Dec 2023 00:14

I thank the authors for making an in-depth effort to address the issues raised by myself and the other two reviewers, which has resulted in largely a new manuscript.

Not all the choices they made the one I would have done, but that is fine, I'm just glad they've had the opportunity to consider our suggestions before beginning data collection/analysis. I don't want to prolong the process un-necessarily, so will refrain from nitpicky comments, but I do have a couple of concerns that affect the core Registered Report design planner shown in Table 1 and thus deserve to be resolved before In Principal Acceptance (IPA):

1) Why are there four analyses specified in section 2.4 (2.4.1-2.4.4) when there are only two hypothesis tests listed in the table? Are 2.4.3 and 2.4.4 "Supporting analyses"? If so, I think these would be better to be described as "Exploratory Analyses" and clearly separated in a different section (I don't think you need to delete them entirely).

Thank you very much again for all your constructive feedback!

Yes, we presented 2.4.3 and 2.4.4 as "supporting analyses". The variance component analysis and beholder index (2.4.3) have been consistently used in studies in the visual domain and we believe they will help deepen our understanding of vocal preferences. Intra-rater agreement (2.4.4) is a "control analysis" to understand, in case we find low inter-rater agreement, if participants were at least self-consistent or not. So, all these analyses are complementary to the whole picture we aim to find.

We understand your point, though, that we need to be more straightforward with confirmatory analyses in the context of a registered report (a task we admit was more challenging than we anticipated, given the largely exploratory character of our study).

As mentioned above, we have followed your suggestion and moved those analyses to an "Exploratory Analyses" section.

2) Several things feel not quite right about the proposed testing of H1 and related power analysis:

a) Why does hypothesis test 1 contain so many different comparisons?

("three pairwise comparisons (paired t-tests, one tailed). Note that we will also compare all styles to each other with a repeated measures ANOVA, potentially followed by 10 pairwise comparisons (paired t-tests, two-tailed))"

What would the interpretation be if only one or two comparisons were supported but not the other(s)? Would this be better broken into, say, three individual predictions? Or modeled in a different way?

(NB: This prediction structure looks similar to one our lab did in our 1st PCI-RR submission for the Hadavi et al. paper I mentioned previously. However, we later updated this after reviewer feedback - perhaps this later update may help give ideas? https://doi.org/10.31234/osf.io/26yg5

Thank you for these suggestions. We recognize we were still mixing exploratory and confirmatory hypotheses. As noted above, we have now moved all exploratory analyses to a clear, designated section (2.4.3).

About the possibility of breaking down our prediction for the singing styles (MM1 lullaby > pop > opera), we did consider breaking it down into three individual predictions as you suggested, but found we had little theoretical support to offer alternative interpretations for all possible outcomes. We ultimately kept the most severe hypothesis test, thinking that non-confirmation of the hypothesis might lead to its refinement or alternatives later on (i.e., following a Popperian approach).

We did, however, make the interpretation of possible outcomes of hypothesis test 1 clearer in the Registered Report Design Planner: only if all three planned (and directional) pairwise comparisons are significant, the results will support our hypothesis of higher shared taste for more “natural”/ universal (lullabies) than for more “artificial” (operatic) kinds of singing, with pop in an intermediary position.

b) Do you really need all those 13 different comparisons (3 pairwise + 1 ANOVA + 10 pairwise)? How is this reflected in the power analysis, which appears to assume 10 comparisons?

("To compare the amount of shared taste across the five vocalization styles, we calibrate our power analysis to have enough power for all 10 pairwise comparisons between styles (and not only the omnibus test). This comes at the cost of conservatively correcting our alpha for ten comparisons ($\alpha = .005$ with Bonferroni correction for multiple comparisons), which, considering our SESOI of $d = 0.5$, necessitates a sample size of 71 participants to achieve power of .9 (paired, two-sided t-test, calculated with the pwr.t.test function from the pwr R package - Champely, 2020).")

(NB: If you focus on fewer comparisons for confirmatory analysis, you might not need as large a sample. Likewise, what is the justification for power of .9? I believe many journals either request minimum power of .95 or .8, so depending on your plans you might also not need so many participants if you are willing to have a lower power (and this could give more flexibility in other analyses).)

Thank you for the suggestions.

Since the journal we are aiming for (Royal Society Open Science) doesn't request any particular value of power, we originally set the value of power .9 as our goal based on the consideration that we were already investing considerable resources in the study, such that it was worth increasing sample size to reach high power, if we could afford that.

As stated above, we recognize we were mixing exploratory and confirmatory analyses in a non-acceptable way, and have now "recalibrated" our power analysis only to our confirmatory hypotheses. Indeed, focusing on fewer comparisons led us to recalculate our necessary sample size. We finally reached a sample size of 60 participants, which ensures power .95 (with three one tailed t-tests, alpha adjusted for three comparisons, and based on our specified SESOI of d=.5).

c) Both H1 and H2 are basically about degree of agreement, right? If so, what is the rationale for using different agreement metrics (MM1 for H1 vs. Krippendorf's alpha for H2)? Wouldn't it be simpler to use a single metric for both tests? Sorry if this is a stupid question, but others will probably have it so would be better to clearly explain.

Thank you for this question, which led to fruitful discussions among us.

As you followed in your role as a reviewer during this rather long process, we had originally planned to address Question 2 with Friedman's test; we then realized that that test would not be sensitive to the differences in the consistency of singer preferences across styles we were interested in, so we brainstormed a different way to track that consistency across styles, and finally had the idea of using Krippendorff's alpha as a measure of "interstyle agreement", as it is perfectly suited for that (i.e., to measure agreement based on mean liking ratings for 22 singers in 5 different styles, which posed as "judges"). Our computations based on simulated and existing data showed that MM1 tends to be higher than Krippendorff's alpha, which leads to the question of which measure would be more appropriate after all. Considering we wanted to (1) connect to experiments from the visual domain, where MM1 is frequently used; (2) profit from MM1's attractive qualities of measuring agreement in relation to the group (whereas Krippendorff's alpha is a wholistic measure for a set of raters and rated items; though note that it is possible to turn Krippendorff's alpha into an individual measure by using a leave-one-out procedure, but we did not want to introduce this new methodological idea here). In the end, all measures of interrater agreement that deserve their name are basically equivalent (as shown in the Figures above), while the important fact is that the mapping of numerical values to verbal, qualitative interpretations is intrinsically arbitrary (i.e., what thresholds should be used to speak of "high" or "low" agreement). As far as we know, there seems to be no natural or universal scale to assess agreement. Hence, in order to homogenize and streamline our analysis plan, we opted to use MM1 in both cases. From our simulations and the analysis of external data, a threshold of .8 seems like a good choice for MM1 and Krippendorff's alpha alike to meaningfully infer "high" agreement. And, as noted above, we still plan to report Krippendorff's alpha as an exploratory analysis, to support methodological discussions on this topic.

d) The new simulation figures I found at https://osf.io/q2sgk?view_only=506d243a6e7a4d3680c81e696ca81025 are great, but would be better to be included in the main manuscript file (perhaps just merge the 2 pdfs into one?). Figs. S2 and S3 help me a lot to understand what is happening to test H2, but Fig. S1 doesn't really shed much light on what is happening to test H1 for me, I'm sorry to say. Could that testing process be made more clear by visualizing simulated data?

Thank you for these suggestions.

We have now simulated new datasets with increasing levels of interrater agreement to illustrate the MM1 analysis of interrater agreement proposed for Question 1 as well (Supplementary Figure S2, which is the same as Figure 2 of the present document, but only for MM1).

We also followed your suggestion of merging the manuscript and the Supplementary Information file for now (we assume this is what you meant; then, for the Stage 2 manuscript, we should separate those files again).

e) Why are the two speaking conditions included at all if they are not the focus of the main tests of H1?

The two speaking conditions were included mainly for hypothesis 2, to test if the same voices will be preferred across all styles, since there is some preliminary evidence of correlations between voice attractiveness ratings for spoken and sung performances (Valentova et al., 2019).

As mentioned before, our study is partially exploratory, and we thought it would be an advancement of the field to provide measures of agreement for contrasting vocalizations in an integrative way (though we lack the theoretical support at the moment to make clear directional predictions for all them). This is why the delineations between exploratory and confirmatory analyses were so unclear.

We have however made an effort to include the speaking styles in hypothesis 1, by making a prediction also for speaking styles, albeit separately from singing styles (i.e., we expect equivalent amounts of agreement for adult- and infant-directed speech performances).

I hope these can help tighten up the manuscript, particularly Table 1, whether that means changing the table or better explaining elsewhere why those decisions were made.

Thank you very much once again for your very helpful feedback!

References:

Valentova, J. V., Tureček, P., Varella, M. A. C., Šebesta, P., Mendes, F. D. C., Pereira, K. J., Kubicová, L., Stolařová, P., & Havlíček, J. (2019). Vocal parameters of speech and singing covary and are related to vocal attractiveness, body measures, and sociosexuality: A cross-cultural study. *Frontiers in Psychology*, *10*, 2029. https://doi.org/10.3389/fpsyg.2019.02029

Review by Christina Vanden Bosch der Nederlanden, 14 Jan 2024 02:24

The authors have done a terrific job of addressing all of the items I and other reviewers have suggested. The restructure of the introduction makes the document very readable and the focus on the main 2 hypotheses, dropping the raft of (interesting!) exploratory analyses also really helps to streamline the proposal. Yet, I still feel one of my comments was not adequately addresses related to the comparison of speech to song. Specificaly, the lack of including speech in hypothesis 1. The predictions seem to only mention song, which dichotomizes the speech-to-song continuum. I understand the authors describe that they do not want to directly compare speech to song because they are considering the whole musi-language continuum, but, by not addressing speech or making any predictions (even if there is little data using your metrics with speech, that indicates to me that this would be novel and useful contribution to the literature!) the authors are treating as a separate category. I would like to see speech added to the analysis for prediction 1 and predictions included for hypotheses 1 and 2. Alternatively, the hypotheses should be split by speech and song, which is logical given that the categories chosen are not equally spaced along a musi-language continuum (that is, none are ambiguous).

Thank you for the thorough and thoughtful responses to all the reviewers comments. I look forward to seeing the data!

Signed,

Christina Vanden Bosch der Nederlanden

Thank you very much for your constructive feedback.

As mentioned in our opening remarks, we have made an effort to include the speech performances in hypothesis 1, and made the prediction (separate from the singing performances, though) that adult- and infant-directed speech should lead to equivalent amounts of shared taste. Of course, this still does not integrate our five vocalization styles in the same way we managed to do in hypothesis 2, and we recognize it still dichotomizes the speech-to-song continuum in a way that we did not intend. However, given the lack of a clear theoretical basis to make our predictions concrete, we hope you will find this solution satisfactory and are of course happy to discuss further this very interesting point.

Review by Christina Krumpholz, 22 Dec 2023 08:57

Dear editors, dear authors,

The planned study and the manuscript have improved a lot and all of my comments have been addressed appropriately. I believe that the planned study profits a lot from the reduced format and the adjusted theoretical framework.

I also had a glance at your provided code which seems to include suited analyses for your planned study.

Therefore, I have no further comments and can with the best of conscious recommend an execution of the study.

Best regards and merry Christmas
Christina Krumpholz

Thank you very much for your positive feedback!