Frankfurt am Main, 30.11.2023

Dear Dr. Chris Chambers, dear reviewers,

We are thankful to all our reviewers for their very constructive feedback. They have allowed us to reflect about our goals and rethink our study design. We greatly restructured our manuscript and believe that we have now reached the high expectations of a registered report.

First of all, we regret immensely that reviewers did not have access to our proposed analyses scripts. Apparently, there was a mistake during submission of the manuscript at the PCI-RR platform (a view-only link to the project folder at osf with all relevant files was supplied, but it seems that the information got "lost" and we hope that this issue will not occur again). We have included a data availability statement with the link to the files in the current version of the manuscript to avoid such problems in the future. All analyses code (also from the previous round of submission) is thus available and can be found in the project folder at osf at the following link: https://osf.io/8k4af/?view_only=506d243a6e7a4d3680c81e696ca81025

Regarding the manuscript itself, after careful consideration, we have reviewed our experimental design and decided to focus on the aesthetic appreciation of contrasting vocalizations, that is, liking ratings, leaving the extensive perceptual ratings of stimuli for a different (more exploratory) study. In addition to reducing the general scope to a crucial and straightforward question, this decision allowed us to include more melodies (three) in the stimulus set, increasing the generalizability of our findings.

We have streamlined our introduction to make our theoretical framework and the rationale behind our experimental design clearer. Additionally, we removed exploratory analyses from the manuscript. This concerns mainly the exploratory analyses about the relationship between acoustic features and liking ratings. However, given the scarcity of empirical research on this topic and the current (limited) state of knowledge, there is still a lot to do in terms of describing and characterizing vocal preferences. This is why we propose important descriptive analyses (variance composition and beholder index – Hönekopp, 2006) that will not be directly used to test predictions, but should help deepen our understanding of individual differences in voice preferences. We illustrate our analysis plan with figures placed in the Supplementary Information (as well as accompanying R markdown files). These are based on the re-analysis of data from a previous study about preferences for pop singing (Bruder et al., 2023), as well as on simulated data. While we hope that reviewers will agree that our decision is reasonable and worthwhile, we are of course open to consider potential alternatives they might propose.

Some major issues were raised by more than one reviewer, so we would like to provide general answers to them first. We would also like to call your attention to an important revision in our analysis plan. After that, we answer to individual points by each reviewer.

- Justification of the stimulus set: In this stimulus set, Brazilian classical singers performed the same melodies in three styles of singing (opera, pop and lullaby) and two styles of speaking (infant- and adult-directed). Stimuli in a language foreign to our German participants (Brazilian Portuguese) have the advantage of having all phonetic characteristics of an existing language, while avoiding semantic confounds (though this issue would in any case be minimized by the fact that the stimulus set is fully-matched, that is, the same material is used for all performances). Note that, after reflecting on reviewers' comments, we have opted to use the version of stimuli with lyrics instead of with a /lu/ sound. Indeed, that allows us to increase the ecological validity of our design. The justification for the chosen performed "styles" was a pragmatic one: recruiting versatile classical singers was a practical way to build a stimulus set with varied and contrasting vocalizations that was fully matched, with all voices performing the same material in all five conditions. We agree with the point made by reviewers that it seems more straightforward to compare speaking and singing directly (e.g., in an infant- versus adult-directed singing/speaking framework), and believe such an exploration of this stimulus set could be an interesting contribution to the current debate between differences and similarities between speech and singing (e.g., Albouy et al., 2023; Ozaki et al., 2022; Sharma et al., 2021). However, as now clearly described in the introduction section, we are interested here in a multidimensional continuum between contrasting human vocalizations - in line with the idea of a speech-music continuum (Phillips, 2023), or with the extension of Steven Brown's musilanguage model (Brown, 2000) proposed by Leongómez et al. (2022, Figure 1b); and in discussing our findings with available literature on (spoken) voice attractiveness. Of course, the relative position of our three different singing styles on such a continuum can be debated, but there is evidence for the existence of meaningful subtypes of vocalizations.

- Revised analysis plan for Question 2 (consistency of average liking ratings by singer across styles): The previous version of this manuscript proposed the Friedman test as a way to compare rankings of singers (built based on average liking ratings) across styles. However, based on simulated data with increasingly more consistent preferences across styles (which were motivated by reviewers' well justified demand for a more thorough power analysis), we realized that the Friedman test was not sensitive/adequate to detect the differences in consistency of preferences across style we are interested in. We have thus revised our plan and now propose to use Krippendorff's alpha to quantify interstyle agreement.

Sincerely,

Camila Bruder,

(on behalf of all authors)

Reviewed by Patrick Savage, 04 May 2023 02:27

I applaud the authors for taking on the challenge of using the Peer Community In Registered Reports (PCI-RR) framework to undertake interesting, largely exploratory, analyses. I find the proposed topic of preferences for speaking and singing voices interesting and valid in principle, though it needs some refinement in terms of theoretical and methodological framing. So in principle I support eventual Recommendation of an improved version via PCI-RR.

That said, I think the current manuscript needs substantial revisions before it meets the standards of a PCI-RR Stage 1 Protocol. In particular, one of the primary goals of RRs is to clearly separate confirmatory and exploratory analysis. Table 1 does this to an extent, but in the main text confirmatory and exploratory analyses are often mixed within the same section or paragraph such that they are hard to distinguish (e.g., Hypothesis 1.1.3 is described as ""not included in Table 1 because it is exploratory", and I think the same might apply to hypotheses predicting liking ratings from acoustic features?). I think all exploratory analyses, variables, etc. should be moved to a separate section and clearly labeled (e.g., "Section 3: Exploratory analyses". I also have some concerns about the generalizability of the stimuli and experimental design, which I suggest the authors consider carefully before deciding whether to continue with the current design or revise.

[NB: In a recent submission from my lab (Hadavi et al., Under review), the PCI-RR recommender actually requested during pre-screening that we completely remove all exploratory analyses for this reason. Personally I think this may sometimes be too drastic and overly limit the ability of authors to use PCi-RR for exploratory research, so I have not recommended it for this case.]

I understand this may be a challenge for the current research, which the authors admit is largely exploratory. I would encourage the PCI-RR Recommender to discuss this issue with the Editorial Board, as the question of how best to use RRs for exploratory research is I believe still an ongoing one without a fixed answer. I strongly support making RRs as flexible as possible to allow for more exploratory work, so my review here is intended to provide constructive suggestions for how to achieve this. I will add that my lab has submitted three manuscripts to PCI-RR on related topics of song/speech/music cognition that are in different stages of the review process, and I encourage the authors to refer to these for ideas/templates for how to reorganize their manuscript/experimental design to make better use of the format (Chiba et al., In Press; Ozaki et al., Accepted In Principle; Hadavi et al., Under Review).

My main suggestions are as follows:

1) Move all details relating to exploratory analyses to a separate, dedicated section (see above)

Thank you for this suggestion and all your helpful comments.

We have in fact removed all exploratory analysis from the manuscript (except for what we now refer to as "supporting analyses", which are descriptive and should help us characterize voice preferences and enrich the discussion of our findings.

2) Add figures visualizing the main confirmatory analyses. *In my experience, it is worth collecting a small amount of pilot data (this can even be just from your three coauthors and/or* *lab members, colleagues) to show proof of principle. This could potentially be combined with the simulations recommended by Lisa De Bruine, although perhaps just the simulations alone might be enough. Even if you only use simulations, I still recommend plotting them in the manuscript.*

Thank you for this suggestion. We have included figures to illustrate our analysis plan in a Supplementary Information file, based on analyses both of previous and of simulated data (and we provide code to run these as well). For instance, we have included the following figures to illustrate simulated data with increasingly consistent preferences across styles, and our proposed analysis to measure this interstyle agreement with Krippendorff's alpha:

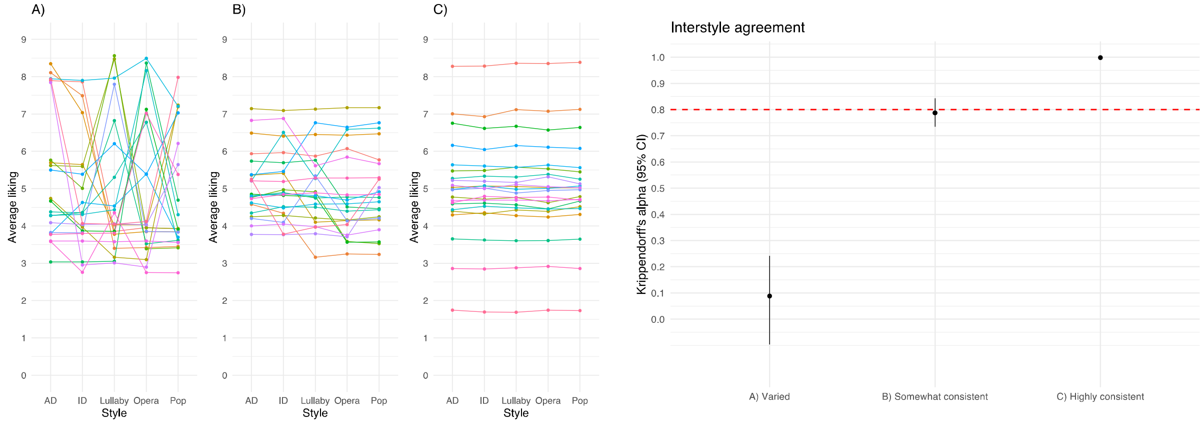

Figure 1: Left: Illustration of simulated data with increasing levels of consistency in preferences for certain singers across the five styles of vocalization. The three simulated scenarios represent: A) varied preferences; B) somewhat consistent preferences; and C) very consistent preferences across styles. Right: Resulting interstyle agreement for the three simulated datasets, as an illustration of proposed analysis for Question 2.

3) Add a "Data/code/stimuli availability statement". *Regarding the simulated data, I note that the authors say they have added R scripts simulating data, but I cannot find those scripts. I recommend uploading them to a repository (e.g., GitHub) and adding a "Data/code/stimuli availability statement" before the reference section incluing this link. I also recommend uploading the full stimulus set here (the manuscript links to a partial stimulus set at https://owncloud.gwdg.de/index.php/s/6IWIvTc828vB77R, but elsewhere says that "[the stimulus dataset, along with details about the validation experiment and acoustic analyses of the stimuli, will be, at the time of publication of this paper, available open access - currently work in progress]". I think RR format best practice would be for the stimulus set to be fixed and open access before receiving In Principle Acceptance (IPA; cf. Chiba et al. In Press for an example).*

Thank you for these suggestions. We have added such a Data/code/stimuli statement to the manuscript.

About the availability of the stimulus set, we have placed the subset of stimuli (three melodies) proposed for this report in the project folder at osf, which is for now still set to "private". We have opted to have the Stage 1 manuscript "embargoed", that is, not openly published after In Principle Acceptance.

Just as a clarification: we are well advanced in the preparation of a manuscript that describes the complete stimulus set (with six melodies), so the whole stimulus set will be published open access soon (which is why we thought we could delay uploading stimuli to the repository for now).

4) Explicitly state how you are correcting for multiple comparisons. *In your response to Lisa De Bruine, you say that the scripts do this, but Table 1 still just shows p<.05 without mentioning any correction.*

To correct for multiple comparisons, we are adjusting p-values with the Holm method. We have addressed that in the revised version of the manuscript (Table 1 and Section 1.3.1).

5) More clearly connect the introduction, hypotheses, stimulus selection, and participant recruitment. *At minimum, I believe PCI-RR requests that submissions use this template including an additional column: "Theory that could be shown wrong by the outcomes" (https://osf.io/sbmx9).*

[NB: I think the version used by the authors may have come from a different website (maybe linked from https://www.cos.io/initiatives/registered-reports?) - I recall having a similar discrepancy in the past and would encourage the PCI-RR Editorial Board to try to standardize this to avoid future confusion.]

Thank you for pointing that out. We have completed the table with that missing column.

While I find the general title topic of speaking/singing voice preferences of great interest, the current discussion of previous literature focuses on hypotheses about the evolution of infant-directed vocalization, but does not make it not clear how the proposed analyses would be interpreted with respect to these hypotheses. For example, it is very difficult to propose predictions that can uniquely falsify theories of lullaby evolution via credible signaling but not social bonding, and vice versa (cf. Savage et al., 2021). The authors propose 5 key conditions, mostly corresponding to infant-directed song, infant-directed speech, adult-directed song, and adult-directed speech, with adult-directed song further divided into operatic and popular. However, the theoretical rationale for dividing into operatic and popular is not clear to me, and even the division between infant-directed and adult-directed is not clear, especially since it seems that both the singer/speaker and the participants listening are probably all adults? Given this, it would seem more natural to focus the experimental design on the singing/speaking contrast, and not worry about infant-directed vocalization. Thus I'd recommend citing and discussing Ozaki et al. and many of the references cited therein (e.g., Albouy et al., 2023; Livingstone & Russo, 2018; Ding et al., 2017; Sharma et al., 2021) regarding singing/speaking contrasts, without so much discusson of infant-directed vocalization. This may also help to refine the confirmatory hypotheses/analyses a bit.

Thank you for all those suggestions. The very good points raised by the reviewer highlight that our general objective and thus our introduction were not clear. In the proposed study, our objective is to investigate aesthetic preferences for the human voice in a wide range of contrasting vocalizations. Inspired by the approach of Vessel et al. (2014, 2018) in the visual domain, we planned to contrapose lullabies, from a more “natural”/universal kind of singing, to operatic performances, a highly “cultivated”, “artificial” kind of singing. By proposing potentially "intermediate" categories or styles of vocalization, we hope to characterize voice preferences with an integrative approach, articulated with the literature on (spoken) voice attractiveness. This led us to the pragmatic decision of recruiting versatile, highly trained female classical singers, to record a fully matched stimulus set, where the same voices produce as many contrasting styles of vocalizations as possible. While we are generally interested in singing vs. speech contrasts and believe this stimulus set would allow for many such interesting comparisons, that is thus not our primary goal. We have now streamlined our introduction to make the theoretical framework and our purposes clearer and added (highly relevant) literature on the speech/music literature in order to provide information to the reader about the usual research interests from which our research departs.

On a related note: who are the 22 females who provided the recordings, and why were they chosen? Are they all trained opera singers (as I would guess from listening to some stimuli)? What languages do they speak? Are they intended to be representative of some broader general population? Should there be some control group(s) to show what effects sex, musical training, language, etc might have? Again, cf. Chiba et al. In Press for examples of selecting stimuli of different types to test hypotheses (in Chiba et al.'s case, low vs. high variance in performer quality; Western classical vs. Japanese folk instruments).

Singers were all Brazilian and speakers of Brazilian Portuguese. They were recruited through the professional connections of the first author, who is a trained classical singer from Brazil. Though (at least in terms of singing ability) these singers are not meant to be representative of a general population of non-singers, we believe that they constitute a representative sample of the population of trained singers able to perform in different styles (and we actually anticipate their speech productions to be representative of a general population). That is to say, we believe their operatic/pop/lullaby singing are not only Brazilian operatic/pop/singing, but generalize beyond singers’ nationality to “singing in general” (instead of something like “Brazilian singing”). In the case of the speech performances, we have no reason to suspect they are different from speech performances of non-singers. We also expect the contrast between infant- and adult-directed vocalizations to generalize beyond singers’ nationality, though of course the language spoken is Brazilian Portuguese.

With the proposed design, we don’t anticipate language influences to be an issue since 1) Brazilian Portuguese should be a foreign language to our German participants; 2) all performances use the same musical and lexical material; and 3) each block of trials will have performances by all 22 singers in one style of vocalization, which should lead participants to focus on the voices themselves.

On the issue of sex and musical training effects, we are collecting data on participants' self-reported gender, sexual orientation and music sophistication (as measured by Gold-MSI - Müllensiefen, 2014). We also plan to ask participants if they recognized the language of the stimuli. However, this information is intended solely for exploratory analyses. These may guide us in identifying potential relationships that could warrant more detailed investigations in later stages of our general research program.

The same goes for participant recuitment: I see the authors have added a statement about this in response to Lisa De Bruine's point, but I didn't see many more details about them in this statement beyond "Participants will be recruited from the participant database of the Max Planck Institute for Empirical Aesthetics's, in Frankfurt, Germany, which consists mostly of lay listeners, with a preponderance of students and retired subjects". Does this database only include adults? (If so, see above regarding whether this is appropriate for testing theories of infant-directed vocalization). Are they all German speakers? Do we need to think about gender? (I suspect male ratings of preference for female vocalization may be quite different from female!)

We have expanded that statement in the manuscript. We also 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). Following the recommendation to limit exploratory proposals in registered reports, we focus on this specific population (to be compared with literature in the visual domain, see Vessel et al., 2018) but plan to suggest, in the discussion section, that follow up studies should extend the investigation of voice preferences to more varied participant samples.

As now clearly stated in the introduction, we focus here on aesthetic preferences of contrasting vocalizations and propose to test adult listeners. Building on the current study and examining the development of such preferences would be extremely interesting (also in relation to current theories of infant-vocalization). We would like to suggest such follow up ideas in the discussion.

Concerning the gender of participants in relation to the gender of the performers, based on current sexual selection accounts of voice attractiveness, one could indeed expect to find gender effects. For instance, Babel et al. (2014) reported that males generally rated fellow males as less attractive than females did, whereas female voices, on the other hand, were rated similarly by female and male participants. However, in the same study they observed that attractiveness ratings by males and females were highly correlated, suggesting the same voices were preferred by both genders. In our case, there are only female voices in the stimulus set, and we are not specifically focusing on examining potential gender effects. Note that the literature on this topic is extensive but particularly mixed, which makes it difficult to even draw a clear hypothesis about gender differences or how they may vary depending on the vocalization style. Nevertheless, we wish to not bias the findings and propose, for the current study, to collect a relatively balanced number of participants from both genders, and to run exploratory analyses (to be reported along with the raw data) on a potential relationship.

Concerning the languages spoken, we will test participants living in Germany and expect that only few of them might understand the language of the stimuli, which is a convenient way to investigate how much they like the voices themselves (without semantic confounds), while using real language stimuli. In any case, we will ask them which languages they speak and if they recognized the language spoken in the stimulus set.

On the issue of language, I was very surprised to hear that all vocalizations only included lexically meaningless "lu" vocables. While this may avoid issues of language confounds, it also doesn't seem to me to be an appropriate proxy for "speech", which by definition uses lexically meaningful words. It sounds like there are also recordings with words as well - my suggestion would be to use those if you have to choose. You could also try running pilot experiments both ways (real lyrics and vocables) to get a sense of feasibility - and perhaps even include both if needed (though I imagine this may be logistically challenging for a long experiment). Cf. Ozaki et al. (Accepted In Principle) for ideas for comparing singing, the same lyrics recited as speech, and conversational speech (which has slightly different acoustic profiles from recited lyrics).

Thank you for your suggestion. After careful thinking, we have reconsidered our choice and plan to use the performances with lyrics instead of /lu/. In the stimulus validation experiments (reported briefly in the current manuscript and extensively in Bruder & Larrouy-Maestri, 2023), the proportion of correct recognition of performances with lyrics was only slightly higher than that of performances with /lu/. It seems that, even though German participants did not understand the language, the phonetic variability of performances with lyrics increased style recognition. In the case of singing/speech preferences, we would expect results to be similar with both /lu/ and lyrics (and hope to address this specific question in a future study) but ultimately decided to privilege the ecological validity of performances (i.e., using performances with "real" foreign language). Note that, as mentioned above, we will ask participants if they recognize, speak, and understand the language of the stimuli.

On a less core but also important note - it sounds like all stimuli are restricted to "only use one of the melody excerpts, the first phrase from “Chove Chuva” (by Brazilian artist Jorge Ben Jor)", which I believe is shown in Fig. 1. This seems like it will pretty dramatically limit the *generalizability of the results to other songs and genres. And why choose a Brazilian piece (in Portuguese?) to test German participants? Consider expanding /changing the stimulus sample.*

As stated earlier, we have reconsidered and decided to expand the number of melodies used to three to increase the generalizability of results.

Concerning the rationale behind choosing to use Brazilian music to test German participants, this approach has the advantage of providing tonal, appealing, pre-existing and human-composed music that is also unknown to participants. And, importantly, the key is that participants will listen to the same material for all styles of vocalization, so they should really focus on how much they like the voices themselves throughout the experiment.

Having said that, it would be very interesting to compare aesthetic responses to these stimuli in populations sampled from different parts of the world, which is one of many possible follow-up studies we plan to suggest in the discussion: there are numerous interesting approaches to examine cultural aspects of certain concepts such as speech/song (e.g., Ozaki et al, 2023) or preferences.

These are all pretty core issues that all may affect the ability to reach meaningful conclusions after collecting and analyzing data. The great thing about RRs is that it is not too late to change this design before you do this! I strongly encourage you to consider my comments here and revise some or all of your experimental design and hypotheses appropriately. (Not saying at all you need to implement all my suggestions, but I do recommend considering them carefully.)

We are very grateful for these important suggestions, which helped us take important decisions regarding data collection and analysis to reach our objectives. The RR process is definitely a constructive and exciting experience.

Minor points:

I also have a few more minor points I'd recommend considering:

-Some statements need references (e.g., “Voice attractiveness has been shown to co-vary with sexually dimorphic traits.”)

-Paragraph beginning: “A first step in this direction was taken in a previous study (Bruder et al., 2021a, 2021b/in preparation)…”: Is it appropriate to rely so heavily on unpublished/in prep. conference presentations when readers don’t have access to them to confirm? I’d suggest not doing this unless there is a written preprint available that people can consult for details if needed.

We agree. The manuscript is now advanced in the review process and the preprint (Bruder et al., 2023) is available to interested readers.

-““pop singing is defined here as singing without any specific type of technique” - I think pop singers may be offended by this, as most pop music does use a variety of genre-specific techniques - could this be phrased differently?

This is also a good point. We did not mean to disregard the skills necessary to perform in pop style. We have rephrased that passage to avoid that meaning (lines 96-99).

-Please explicitly specify independent variable(s) (vocalization type?) and dependent variable(s) (liking rating?)

Sorry for the lack of clarity in the previous version. Liking ratings are the dependent variable and the "styles of vocalization" is the independent variable. This is now clearly stated in the revised manuscript.

-“performed one fifth higher as pop and lullaby stimuli” - what does this mean and why was it done?

Performances in operatic singing were recorded with higher pitch than pop and lullaby performances. This was done to produce naturalistic performances and to keep singers comfortable, since operatic singing typically has higher pitch than pop and lullaby singing. This choice complicated our acoustic comparisons of stimuli (in the sense that it limited insights about differential production mechanisms) and likely impacted style recognizability by listeners, as we addressed in detail in our manuscript focusing on the versatility of these classical singers (Bruder & Larrouy-Maestri, 2023). For the current study, however, this should not pose any issue. Indeed, we are interested in how much participants agree in terms of which voices they prefer and all performances in operatic style will be presented in the same block of trials, thus avoiding a direct comparison with other performances.

I hope these suggestions are constructive, and wish you good luck in trying to appropriately revise the project!

Thank you very much! Indeed, they were very helpful and constructive.

Signed,

Patrick Savage

PS For transparency, I wish to disclose that two of these three authors are coauthors with me on a mega-collaboration with over 70 coauthors on the topic of speech and song that has received In Principle Acceptance from PCI-RR (Ozaki et al. Accepted In Principle; names bolded below). I have not otherwise collaborated with or otherwise have conflicts of interest with *any of the authors. I confirmed with PCI-RR before accepting the review that such mega-collaboration coauthorship does not disqualify me from serving as a peer reviewer.*

PPS: The linked PDF did not appear to incorporate the changes from the previous revision - fortunately I downloaded the tracked change file which did! But in future please try to ensure the revised version is correctly uploaded.

We are very sorry about that. We will make sure that does not happen again.

References:

Chiba, G., Ozaki, Y., Fujii, S., & Savage, P. E. (In Press). Sight vs. sound judgments of music performance depend on relative performer quality: Cross-cultural evidence from classical piano and Tsugaru shamisen competitions [Stage 2 Registered Report]. Collabra: Psychology. Preprint: https://doi.org/10.31234/osf.io/xky4j (Peer Community In Registered Reports editorial recommendation and peer review: https://doi.org/10.24072/pci.rr.100351)

Hadavi, S., Kuroda, J., Shimozono, T., & Savage, P. E. (Under Review). Cross-cultural relationships between music, emotion, and visual imagery: A comparative study of Iran, Canada, and Japan [Stage 1 Registered Report]. PsyArXiv preprint: https://doi.org/10.31234/osf.io/26yg5

*Ozaki, Y., Tierney, A., Pfordresher, P. Q., McBride, J., Benetos, E., Proutskouva, P., Chiba, G., Liu, F., Jacoby, N., Purdy, S. C., Opondo, P., Fitch, W. T., Rocamora, M., Thorne, R., Nweke, F., Sadaphal, D., Sadaphal, P., Hadavi, S., Fujii, S., Choo, S., Naruse, M., Ehara, U., Sy, L., Parselelo, M. L., Anglada-Tort, M., Hansen, N. Chr., Haiduk, F., Færøvik, U., Magalhães, V., Krzyżanowski, W., Shcherbakova, O., Hereld, D., Barbosa, B. S., Varella M. A. C., van Tongeren, M., Dessiatnitchenko, P., Zar Zar, S., El Kahla, I., Farwaneh, S., Muslu, O., Troy, J., Lomsadze, T., Kurdova, D., Tsope, C., Fredriksson, D., Arabadjiev, A., Sarbah, J. P., Arhine, A., Ó Meachair, T., Silva-Zurita, J., Soto-Silva, I., Maripil, J., Millalonco, N. E. M., Ambrazevičius, R., Loui, P., Ravignani, A., Jadoul, Y., **Larrouy-Maestri, P., Bruder, C.**, Aranariutheri, M. I., Teyxokawa, T. P., Kuikuro, K., Matis, T. U. P., Natsitsabui, R., Irurtzun, A., Sagarzazu, N. B., Raviv, L., Zeng, M., Varnosfaderani, S. D., Gómez-Cañón, J. S., Kolff, K., Vanden Bosch der Nederlanden, C., Chhatwal, M., David R. M., I Putu Gede Setiawan, Lekakul, G., Borsan, V. N., Nguqu, N., & **Savage, P. E.** (Accepted In Principle). Globally, songs are slower, higher, and use more stable pitches than speech [Stage 2 Registered Report]. Peer Community In Registered Reports. Preprint: https://doi.org/10.31234/osf.io/jr9x7 ((Peer Community In Registered Reports editorial recommendation and peer review: https://rr.peercommunityin.org/articles/rec?id=316)*

Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., & Fitch, W. T. (2021). Authors' response: Toward inclusive theories of the evolution of musicality. Behavioral and Brain Sciences, 44(e121), 132–140. https://doi.org/10.1017/S0140525X21000042

****

Anonymous review

Overall summary:

The paper proposes to examine how vocal aesthetic preferences vary across specific speaking and singing styles. The main novelty of the current proposal is the inclusion of music or song aesthetics as the introduction describes a large literature related to vocal attractiveness in speech. The proposal will examine how much agreement there is among raters in their liking of primarily sung materials, with a specific examination of whether agreement about particular performers differs depending on the genre or style. Interesting metrics from the visual literature, MM1, will be used to assess agreement alongside more traditional metrics of inter-rater reliability (ICC). The proposed study is novel and has the potential to contribute meaningfully to the field of empirical aesthetics, but there are several items that should be addressed before moving to IPA. First, the introduction is a bit hard to follow in that the reader could arrive at multiple different main study questions from the literature review. That is, since the novelty of the current proposal appears to be from a lack of studies related to song, I assumed that the main comparison would be speech vs. song, but that is not the case. The introduction should be streamlined and written to clearly set up the main goals and gaps in the literature that the proposed work is designed to address (e.g., why the test-retest reliability? Why these styles? Why these acoustic features?). The second issue is that the proposed study has a lot of additional metrics (all the perceptual factors) that the participant pool will need to complete that appear to be ancillary to the main question about liking. Instead this study design feels more like a validation of the stimulus set and that the liking question was tacked on as an afterthought. That is a completely normal approach, but for a pre-registration, I feel the study design should be more clean and deliberate so as to answer a specific question without the unintended influence of multiple additional factors built into the study design for exploratory analyses. Finally, I believe the power analysis could be more carefully down with reference to previously published work.

Thank you for all your constructive feedback. As you will see in the list below and in the revised version of the manuscript, we addressed (or clarified) every point very carefully:

We realized that the main research question was unclear. We are indeed not aiming at a direct comparison of speech/song (as in, for instance, Albouy et al., 2023; Ozaki et al., 2022; Sharma et al., 2021), but focus on how shared are preferences across a set of contrasting vocal performances. We streamlined our introduction to make the theoretical framework clearer and to properly set up our goals and justify how we plan to achieve them. We aim to: characterize aesthetic preferences for the human voice in an integrative way; test the prediction that there will be more shared taste for more “natural” (lullabies) than cultivated/”artificial” (operatic) types of singing; and test how consistent are average preferences for some voices across contrasting styles of vocalization.

Importantly, we revised our experimental design. The extensive perceptual ratings were discarded in order to focus on the liking ratings (and allow the use of more melodies, please see previous comment about generalization). Also, we revised our analysis plan and made sure to expose it in a clear way. As mentioned above, we now propose a different approach to answer

Question 2: based on average liking ratings by singer for each style of vocalization, we propose to quantify the consistency of average preferences across styles by measuring the “interstyle agreement” with Krippendorff’s alpha.

Also, we revised our power analysis. It is now informed by previous data: we calculated MM1 for previous data about preferences for pop singing (Bruder et al., 2023), and defined a difference of .1 in the computed MM1 value (based on the observed average value of 0.44, SD = 0.2) as our smallest effect size of interest (SESOI). This decision governed our sample size estimation (now increased from 45 to 71 participants).

General comments:
Throughout the manuscript there is a differentiation between perceptual features and acoustic features, but this is not defined. For instance, breathiness, tempo, and timbre are described as perceptual features, but they can be just as easily characterized as acoustic features extracted from onset rates, spectral cues, harmonicity from Praat or MIR toolbox-like metrics. Can the authors please define their meaning at the outset so the reader can assess what factors (human vs. algorithm based metrics, musical vs. speech-based metrics, etc) they are differentiating when they describe the contribution of these features.

You are right about the lack of definition of acoustic and perceptual features in the text. We were indeed referring to computationally-extracted “acoustic” versus participant-based “perceptual” ratings. Ideally these two sets of features (i.e., acoustic and perceptual) are highly related to each other but, based on previous work (see preprint – Bruder et al., 2023), that relationship is not that simple: we found that even though average perceptual ratings correlated with acoustic measurements in the expected direction, interrater agreement on the perceptual scales was very low; and we could predict very little variance of liking ratings based on acoustic measurements, but around 43% based on perceptual ratings – so there seem to be highly variable individual differences in how participants perceive singing performances, both in general (i.e., perceptual ratings) and in terms of aesthetic appreciation (i.e., liking ratings). Note that for this revised proposal, considering we are no longer collecting perceptual ratings, we believe this issue might be out of topic and thus confusing to the reader, so we do not really address it in the revised manuscript.

Abstract:
What is the relevance of vocal attractiveness to the bigger picture of sociobiological signals?

The “type” of vocalization could be interpreted in many ways, but I think you mainly mean speaking vs. singing, although I do see your different types of vocalizations, too. Perhaps your main proposal sentence could read something like: “For instance: why do we like some voices more than others? Does our liking of voices differ depending on whether a person is speaking or *singing? Do some voices sound better in some contexts like singing a pop song but not singing an operatic aria?*

Thank you for these ideas. We are focusing on the contrasting categories of vocalization rather than on the speaking vs singing contrast. As now clearly developed in the revised version, we propose to make a parallel with findings from the visual domain (e.g., Vessel et al., 2014, 2018) and contrapose styles which are increasingly "cultivated" (though, weather findings are similar to the ones in the visual domain or not, we believe that the results will advance current knowledge about voice appreciation in an integrative way).

What is the difference between perceptual and acoustic features? Do you mean features based on participant's ratings of acoustic characteristics? Do you mean preferences here not perceptual features?

As mentioned above, we were referring to computationally-extracted "acoustic" versus participant-based "perceptual" ratings, as in the contraposition made in our previous work (Bruder et al., 2023), where we found that plain acoustic features explained very little variance of liking ratings, while (individuals' ratings of) perceptual features explained around 43% of variance. In the present RR, we use liking ratings to learn about the aesthetic appeal of voices, and use the term "preference" quite generally, as in: higher liking ratings for a certain singer indicate she was "preferred" (even though we don't use a pairwise comparison design).

Introduction:
I appreciate the different functional roles, esp. described for ID speech, but it would be good to characterize the role of AD speech as well – I see that you write "Beyond supporting interpersonal communication and conveying semantic information" which I assume is about AD speech, but it would be nice to flesh that out in a sentence or two and then continue with your shift in narrative toward it being a sociobiological signal. This section could do a little more for the reader by summarizing the similarities and differences between music and language.

There needs to be a better segue to the proposed work that motivates why the question is important to study and relevant to human communication. I believe it is important, but more of a motivation is needed to help understand your hypotheses and the context for your proposed study. For instance, why does it matter if vocal attractiveness differs for the same speaker speaking vs. singing? Or is the question more about what features predict attractiveness? Or whether vocal attractiveness is even important for some modalities vs. others?

You are absolutely right when pointing out our manuscript was lacking a clear direction. Given the initial stage of empirical investigations about (singing) voice preferences, all of the questions you mentioned need to be addressed (and interest us immensely!). But, as stated above, this proposal does not aim to contrast speech and song. We are sorry for the misunderstanding and have clarified our purposes in the revised manuscript.

Also, after consideration, we have decided to leave the question of predicting attractiveness from features aside for a different study and to focus on the general characterization of voice preferences, making the parallel with findings on the visual domain and quantifying the amount of shared taste across contrasting styles of vocalization.

Vocal attractiveness and sexually dimorphic vocal features: In regards to harmonic-to-noise ratio (called harmonicity in Praat), the cited work is relevant and makes sense, but it a little older. Especially given that there is a strong trend, especially in female voices, to use vocal fry which would have, I assume, a lot more noise in the signal than harmonic information. This was originally associated with California valley girl speaking style (i.e., cool!) but has now pervaded all of north America and likely beyond. It's work having a sentence about this in this section to update the literature review.

Thank you for this suggestion. Since we removed all exploratory analyses involving acoustic features and restructured our introduction, our literature review does not cover the acoustic bases of voice attractiveness in detail, so we in fact removed that older paper from the introduction.

Inter-rater agreement is low (how low?) can you contextualize this for the reader?

The inter-rater agreement was really low (the average across 10 perceptual scales Krippendorf's alpha was .15; we found very similar numbers with ICC2). We have added this information to the manuscript in order to provide some context to the reader.

RE: Visual data using MM1 – I am having a hard time understanding this section: "They argue that the behavioral relevance of naturally occurring types of stimuli such as landscapes and human faces results in information processing, and hence aesthetic experience, that is highly conserved across individuals."

What is meant by behavioural relevance?

So is the use of MM1 about being integrative or about comparing with vision or about assessing behavioural relevance? Clarity here will help guide what hypotheses should be or what hypotheses are expected by the reader.

We are sorry for the lack of clarity in the section. We use MM1 to allow for comparison with the visual domain, which serves as a theoretical framework to make predictions. In this context, it has been proposed that the higher shared taste found for more natural kinds of stimuli can be explained by their higher and more uniform (across individuals) behavioral relevance, which would lead to shared associations and, ultimately, shared taste (Vessel et al., 2018). Behavioral relevance can be understood here with the general meaning of how important a stimulus is to an individual, not only in terms of sexual selection (e.g., face preferences as adaptations for mate choice, signaling mate quality – e.g.,Rhodes, 2006), but also in how learned associations with those stimuli may influence future choices and behavior. Thus, in the case of natural kinds of stimuli, these experiences and resulting associations tend to be similar for most individuals, which should lead to more shared taste. On the other hand, in the case of more cultivated/"artificial" kinds of stimuli, the lack of uniform behavioral relevance for most individuals (and hence lack of similar experiences and shared associations) allows for the expression of idiosyncratic preferences.

In the case of our five categories of vocalization, performances are all natural (in the sense that they are not synthetic); and they are all clearly behaviorally relevant, though probably not uniformly so. This leads us to expect differences in the proportion of shared taste for lullabies, as a more "natural"/universal kind of singing, and lower shared taste for operatic singing, as a more "cultivated"/very specific kind of singing, with pop singing in an intermediary position.

In the revised manuscript, we have more clearly explained our integrative approach and defined the notion of behavioral relevance, and believe that that will help the reader better understand the hypotheses actually tested with the proposed design.

Questions & Hypotheses (Table 1)

I think the main questions are interesting, but it feels like some of the simple effects are missing. For instance, since there are 2 modalities (speech vs. song) and within each modality 2 to 3 sub-types (song: opera, pop, lullaby; speech: AD vs ID), I imagine you would want to examine whether the same speakers are preferred across sung and spoken stimuli and then drill down into whether that interacted with "styles." However, I do see that the spoken and sung styles are really not comparable between modalities. For instance, song as ID and AD registers as well and that would be a better comparison of differences between speech and song. Or speech has casual conversational styles, conversing with strangers, delivering a speech, or acting on stage, which could be more comparable to the categories chosen for song. So if this is the rationale for not looking at speech and song separately because the subgroups are 1) not balanced and 2) not directly comparable to one another, then that is fine, but I think it's worth noting this rationale somewhere so the reader understands the rationale for the study design.

We agree the rationale for the study design was not clear enough and hope to have satisfactorily solved that now.

Hypotheses 1.1.1 only includes song, but the explanation for it includes speech but does not include a directional hypothesis. Speech should be included in the analyses and some hypotheses should be made either about speech sounds alone or in comparison of speech to song, otherwise I am not sure why the speech stimuli are included.

The speech stimuli are included to explore voice preferences with an integrative approach. We only included two styles of speech for practical concerns (i.e., related to the complexity of recording and handling so many performances) and this may indeed be seen as a limitation of our approach. In fact, we would be very curious to know how much voice preferences vary (or not) across other substyles of speech (e.g., for the subtypes of casual conversational styles, conversing with strangers, delivering a speech, or acting on stage you mentioned in the previous question) and hope to address this issue (or that someone would do) in the future. Regarding the lack of a directional hypothesis for the speech stimuli, the literature on this topic does not allow to make predictions. Studies of voice attractiveness rarely report interrater agreement, so individual differences in aesthetic evaluation of voice (for speech as well as singing) remains to be examined.

1.1.2 – in the table these are referred to as rankings, but they are ratings. Describing them as rankings made me think that participants might be doing a ranking task instead of the researchers using the average ratings to effectively rank the performers based on participants ratings.

Indeed, we were referring to rankings built based on average ratings, and we could have been clearer on that. As you can see in the revised version, we revised our analysis plan to assess interstyle agreement with Krippendorff’s alpha instead of comparing rankings with Friedman test. Therefore, there is no ranking anymore and this section had been modified accordingly.

What is the duration of time between test and re-test? I don’t see this in the experimental procedure. What have other studies done? 2 weeks? 2 months? Back-to-back days does not seem long enough to truly examine consistency.

We propose a maximum of 14 days between testing sessions, having the two sessions one week apart from each other when possible. The rationale behind this comes from the consideration that none of the two aspects that should be considered when testing reliability with the test-retest method (the possibility of learning, carry-over, or recall effects; and the possibility of a change in status of the measured trait between sessions - Allen & Yen, 1979) is of particular concern in our paradigm: since we have so many stimuli (330), we estimate the possibility of participants remembering their answers from one session to the next to be negligible. And, considering that music abilities and engagement seem to be relatively stable among adults (Müllensiefen et al., 2014) we have no reason to expect participants’ general preferences for voices (or melodies) to change much across time. Thus, we based our definition of the test-retest interval mainly on pragmatic constraints related to data collection.

We have now made the planned interval for retest (and the rationale behind it) clear in the manuscript (Session 2.3, lines 362-376).
We are not sure what you meant with “Back-to-back days does not seem long enough to truly examine consistency”, but we would argue that a high test-retest correlation in liking ratings between the two testing sessions (preferably seven days, but up to 14 days apart) is indeed a good measure of how self-consistent participants are in their preferences. In case you meant that two following days would be too little: we would argue that it is highly unlikely that participants can remember ratings given on a previous day but we propose longer between-sessions time (just in case).

1.1.3 Should include what perceptual and acoustic features you’ll use to predict liking ratings and include what other factors you’ll include, etc.

You are right, this information was missing in the last version of the manuscript. Following reviewers' suggestion, we removed the exploratory aspect of this research. As a consequence, In the reviewed design, perceptual and acoustic features no longer apply.

1.2.1 How did you calculate f? In g-power they have you determine f directly from eta-squared or from variances. Please justify, using the previous papers cited, why you expect an effect of this size and how that relates to the previously found effect sizes (eta-squared). For 1B, what estimates does g-power (z tests section) give you for finding a high correlation between your to dependent correlation coefficients?

As mentioned previously, we have revised our power analysis, which is now informed by previous data about preferences for pop singing (Bruder et al., 2023). We based our sample size estimation on a smallest effect size of interest (SESOI) of .1 in the computed MM1 value (based on the observed average value of 0.44, SD = 0.2), and concluded that, to achieve power of .9 (with two-tailed t-tests and adjusting alpha to 0.005 with Bonferroni correction for all possible 10 pairwise comparisons between styles), we need 71 participants (thus increasing our initially planned sample of 45 participants).
We have conducted this analysis in R and provide R markdown scripts documenting it.
We have justified these choices in our revised analysis plan (Section 1.3.1).

1.2.2 I am curious why the Friedman test is warranted? Is it predicted that the residuals will not be normally distributed in the ANOVA? It seems like the mean rating averaged by performer should be fine in an ANOVA unless there are different number of trials per performer? The ANOVA would be a X (Number of performers) by style (n=5) repeated measures analysis as there will be 5 columns per performer, correct?

Thank you for your comment. The Friedman test was indeed not the best approach to our needs. We proposed the Friedman test as a way to compare the rankings (built based on average ratings) of singers in different styles, that is, to test if some voices are consistently “better”/preferred across styles. That finding would support (or not) the idea that both singing and speech vocalizations work as “backup signals”, conveying the same information about individual fitness (Valentova et al., 2019). However, as mentioned above, we have revised our analysis plan because, based on simulated data, we realized that the Friedman test (when conducted as we originally proposed it) is not sensitive to the differences in preferences across styles we are interested in: in our simulations, both data with highly consistent and with very varied preferences across styles led to non-significant differences when styles were compared with the Friedman test (please see our accompanying R scripts for a demonstration of this, and Section 1.3.2 and Supplementary Figure S3 for an illustration of our alternatively proposed approach to measure interstyle agreement (that is, how consistent are average ratings across styles) with Krippendorff’s alpha.

You are correct that an ANOVA as you suggested would be a 22 (singers) by 5 (style) repeated measures analysis. This analysis would allow to compare average liking ratings between styles and show us if participants liked one style more than others. Though that is a highly relevant question, this does not specifically address our main question about the shared appreciation of specific voices. Note that the data will be available and such a follow up analysis will thus be possible.

1.2.3 I know that these are reported in the spirit of explorational analyses, but if they are reported here then they are pre-registered and so they ought to have more detail, I think. If they are truly explorational then perhaps they can be left off of the pre-registration? Given that the authors have extensive experience doing this sort of analysis based off of previous cited work, it seems as though the authors could do some clear pre-registration of a simplified version of their LMEs so that at least some part of this analysis could be evaluated at Stage 1. As such, a power analysis of some sort is warranted here to ensure that the authors have enough power to do even their exploratory analyses.

Also, since the paper is about whether some performers consistently are liked more than others, then perhaps performer should be included as a variable in the LME? Perhaps, if you have enough participants, including speech and song in one model and coding for it would be informative. This would likely require a lot of interaction terms to determine if a feature was useful for song but not speech, for example, so it can get unwieldy (and not converge) quickly, but it would be a stronger way to illustrate that a features usefulness changed depending on the style or modality than two separate models.

After careful consideration, we still believe that the proposed exploratory analyses are very “tempting”, but we also fully agree that they should be properly planned (and backed by power analysis) to be kept in the registered report. We thus decided to remove these, along with the detailed collection of perceptual ratings.

Participants. “students and retired subjects” do you think you should restrict to one age group or include some sort of coarse age grouping in your study? It seems very likely that there will be generational differences in aesthetic preferences for voices.

That is a good point. Age is indeed a variable that has been shown to interact with music perception (e.g., Fischinger et al., 2020) and preferences (e.g., Hird & North, 2021), and one can imagine that there could be generational differences in aesthetic preferences for voices. However, we do not have a specific hypothesis about that at the moment.

We opted for examining a sample as large and varied as possible within certain practical constraints. Through the database of the MPI for Empirical Aesthetics, we have access to a convenience sample that has the advantage of not being composed only by psychology students (as is commonly the case in studies like ours), and where all age groups are represented. For instance, in our last lab study, participants’ age was on average 36.8 years (SD = 16.1, range: between 22 and 75 years old). But, as mentioned by the reviewer and acknowledged in Section 2.1, several registered participants are students and retired individuals, which is why we already acknowledge this as a potential bias. We also acknowledge that this convenience sample shares the generalizability limitations of most studies sampling from “WEIRD” populations (White, Educated, Industrialized, Rich, and Democratic - Henrich et al, 2010).

That being said, considering that generational differences in aesthetic preferences for voices are not the main objective of our study, we will keep the reviewers' comment in mind and address this point in exploratory analyses.

Stimuli. How will you account for F0 differences that were requested for singers of pop and lullaby in your 3rd hypothesized LME? Same for loudness. You will need to be careful about this if concluding that F0 or loudness predicted liking, for example.

That is also a good point. In this revised design, we are no longer focusing on predicting liking from stimuli features, so we avoid these concerns (for now!) and will definitely take them into account in the future.

“This leads to 110 performances (by 22 singers, each performing three styles of singing and two styles of speaking).” I see why you decided to have the same melody for all singers being judged (apples to apples), but, in order to make strong claims about style and features of that style, it seems like you’d want to include at least one other melody. At the very least, this might make the task a little more enjoyable (and reliable) for participants, given that 22 singers is quite a big number to move through. I know you want variability in performers so keeping the number at 22 is understandable given your hypotheses, but it is worth considering making the study a bit longer for generalizations’ sake.

This major concern was also raised by other reviewers and led us to reconsider our experimental design. We now include three melodies instead of only one, which is now possible since we decided not to collect the planned 10 perceptual rating scales. We fully agree with the reviewer's suggestion. This allows for generalization of findings and probably enhances participants’ enjoyment of the task (probably enhancing reliability as well).

Acoustic and Perceptual analyses. I see here how you have grouped acoustic and perceptual. Basically, perceptual are provided by participants and acoustics are not. But it’s hard to say that the perceptual features are not the same as or highly correlated with acoustic features. For instance, energy – as an acoustic measure, should be highly correlated with perceptual loudness or F0 calculated by Praat should be highly correlated with participants high-low ratings. How will you enter variables into your LME models? What if there is significant correlation among predictors – will you drop the least correlated? Will you compare models with all correlated acoustic features and all correlated perceptual features? Will you compare models that have a mix of perceptual and acoustic features but only those that are not highly correlated with one another? You have a lot of great variables here, but they are largely overlapping so it makes it hard to understand if the story you’re trying to weave in this case is about the failure of music information retrieval techniques to pick up on the features that real human listeners use, or something else. And the selection of these variables should be justified (for instance why perceptual and acoustic measure of pitch?)

You are right on all raised points. In the reviewed experimental design, these concerns no longer apply, but we would like to clarify our reasoning and hope to adequately answer to the reviewer's remarks.

Though it was not clear enough in the previous manuscript, we were building on previous findings (now properly described in the preprint, Bruder et al., 2023). In this study, we found very little prediction of liking based on “acoustic features” (i.e., computationally-extracted descriptions of the audio signal) but some prediction based on “perceptual features” (i.e., based on participant’s ratings of perceptual attributes of the singing performances). To reach this conclusion, we built separate “acoustic” and “perceptual” models, which achieved very low and moderate prediction, respectively. We also showed that, even though interrater agreement on all scales was very low, average perceptual ratings (across participants) correlated in the expected direction with the correspondent relevant acoustic measurements (e.g., average perceptual ratings of loudness correlated with rms energy; average pitch accuracy ratings correlated with estimates of pitch interval deviation, which were based on Praat's F0 estimates; etc). Concretely, there is definitely a relation between acoustic and perceptual features, as rightly pointed out by the reviewer, but these two types of features are not perfectly matching. Therefore, all the raised points make a lot of sense and were actually on our mind, even though they were not specified clearly enough in the manuscript. We are actually planning to delve into this topic at the occasion of future work (since it is out of the scope of the RR) and would be of course happy to discuss this issue further with interested researchers/labs.

Liking ratings are part of this very large set of perceptual ratings. Is there a reason that each person needs to rate these perceptual features? If your study is truly interested in liking ratings, then it feels like these should be two separate studies or that liking should be asked first so that the large list of features participants need to rate does not bias their liking rating. Further, I wonder how having these perceptual features drives liking ratings on subsequent trials. For instance, a rater might intuitively think that diction is crucial to a good performance and then decides to apply that to all sung stimuli, but would not have considered diction had they not been explicitly asked about it during the study. I worry that all of these perceptual ratings will alter participants liking ratings. It also seems to me that these perceptual ratings do not need to be obtained on a per person basis (that is, do you plan to use perceptual ratings from a given person to predict liking or an average from all participants?).

As stated in the previous answer, we found in our previous study (Bruder et al., 2023) that interrater agreement in perceptual ratings of singing was very low, which is why we were initially planning to collect perceptual ratings from all participants in this report. This plan has now changed, since we have decided to focus only on liking ratings (as stated above).

However, we would like to note that we agree that the different perceptual scales may make participants more aware of certain aspects of the performances and thus influence their liking behavior. We are currently preparing a specific experiment to test this and would be happy to share the results when the study is completed.

Procedure, cont'd
Blocks grouped by style – while I agree this is a valid approach, I wonder if the blocking itself might alter the predicted results. Specifically, blocking by style may encourage raters to adopt a set of features for that specific style or genre, whereas varying the style and completely randomizing blocks may encourage people to attend to performer-specific features that are aesthetically pleasing. If you decide to pare down the task to only liking, it could be interesting to run the study with both randomized and blocked presentation of trials perhaps for different

melodies (CH vs. NN) or between subjects to examine this methodologically important difference.

Thank you for raising this very relevant point. We also suppose blocking by style will leave participants free to adopt different strategies for rating each style. In the present case, we believe that it is not a problem per se since we aim to examine agreement within styles. Though it has not been specifically examined in the case of singing yet, we can refer to literature in the visual domain. For instance, we can think of the study by Vessel and Rubin (2010), in which the issue of presenting different categories of stimuli in a blocked versus intermixed design was very clear: when different categories (in the case, real-world versus abstract images) were presented separately, they found high interrater agreement for real-world images (arguably driven by shared semantic knowledge about stimuli) and low interrater agreement for abstract images. When categories were intermixed (i.e., in approximately half of trials participants had to compare real-world images with abstract images), agreement for real-world images dropped to values near the level found for abstract images. The authors argued that presenting stimuli of different categories intermixed likely forced participants to use the same common strategy of deemphasizing meaning to respond to all stimuli, thus basing their decisions more on the visual aspects of the images.

In our framework, considering we are characterizing singing voice preferences for different styles, it seems wiser to allow for such differences to emerge. If we present different styles intermixed, we would likely force participants to use a common strategy to deal with all of them - which also seems interesting and worthwhile, but probably as a second step.

Data analyses. This MM1 metric seems interesting and reminds me a lot of jack-knifing techniques for understanding the contribution of that particular rater (or item) to the mean. However, each participant will have a single R value, correct? If so, then I am not sure I understand the next sentence, pasted for clarity below. Wouldn't the mean of the z-score be 0 and then transforming them back to and r-value would leave them unchanged? I could be missing a step here! I am glad you're including the ICC as this seems to be a pretty standard metric for assessing agreement across raters.

"The across-observer average MM1 score is computed by 1) transforming individual r-values to z values, 2) computing a mean, and 3) transforming that score back to an r-value for easier interpretability."

As well understood by the reviewer, each participant will have a single r value, corresponding to how much his or her ratings correlated with the average rating of all other (N-1) participants. To avoid the issue that averaging raw correlations produces biased results (Corey et al., 1998), we first convert individual r values to Fisher z-scores; average across all participants; and convert back to an r score for ease of interpretation. We are following the procedure proposed by Vessel et al., (2018).

We have improved this section in the manuscript (Section 2.4.1) and included analysis code, which should clear any doubts.

At the end of a proposal I would expect some sort of impact statement about what the predicted results would contribute to the field or what follow-up studies it would spur. I am not sure if this is a typical section for PCI-RR, but it seems important to close out the proposal with this information.

We have included remarks on the significance of the proposal in the end of the introduction, after we specify our goals.

****

Reviewed by Christina Krumpholz, 16 May 2023 09:46

Dear authors, dear editors,

In the following, I'm reviewing the registered report "Voice preferences across contrasting singing and speaking styles" by Bruder, Frieler & Larrouy-Maestri. Generally, I think this is a well-planned study which can extend previous knowledge and lead to interesting findings! However, I have some suggestions that should be taken into consideration before conducting the study.

1. Major: In general, I'm missing a proper theoretical framework for this study. As this is generally a problem in psychology, I'm pointing it out here hoping that you will put more emphasis on this and also consider it in your discussion when interpreting results again!

Thank you for your helpful suggestion. We have restructured our introduction to make our theoretical framework clearer and discuss it properly later in the manuscript. We also acknowledge that, given the scarcity of empirical research on singing voice preferences, this framework is somewhat “adjacent” and “borrowed” from neighboring disciplines.

2. *Major: The introduction would profit a lot from a different structure. Currently, starting with "infant speaking" and then just briefly touching "general speaking" is misleading - I would expect a study that is much more focussed on infant vs. adult talking, which is not the focus of the study. Maybe you can think of a more general introduction that mentions infant talking as one subpoint? This would also better lead to your paragraph of the describing the present study and make the research interest more pronounced.*

Thank you, we agree the introduction needed restructuring. Our literature review is now better suited to our research question, and clearly presents the theoretical framework supporting our study.

3. *Major: On page 2, research questions are defined: "But why do we like some voices more than others? How does our enjoyment of voices vary across diverse types of vocalizations?" - I think it could be formulated more precisely, so that it's clear that you're looking at both shared and individual taste here.*

Thank you for raising this point, we rephrased this section to make it more precise. Concretely, we kept the very general question on the roots of voice preferences (“why do we like some voices more than others?”), but changed the second question to clearly show we mean to quantify shared taste to characterize these preferences (“How much do people agree in their voice preferences across different types of vocalization?”). We avoided the expression “shared taste” for this second paragraph in the introduction, but presented that formally later on.

4. *Major: I agree with what Lisa DeBruine has mentioned before, a data simulation would be advantageous - however, I don't see any code? Would it be possible to send me your planned analysis code?*

We are very sorry about the mistake in the submission that prevented reviewers’ access to our analysis code. The link to our project folder is properly included in a Data/code/stimuli availability statement, and both the old and the updated code are available there. If you prefer that we send to you directly, we will be happy to do it.

5. *About the sample sizes. For 1.2.1. and 1.2.2.: Why are you running your sample size calculations with an expected small to moderate effect size? What is your rationale behind that? SESOI or previous findings?*

We have revisited thoroughly our power analysis based on previous findings. We calculated MM1 for the liking ratings of pop singing described in the now available preprint (Bruder et al., 2023), and determined that our smallest effect size of interest would be of a 0.1 difference in MM1 values between styles, which corresponds to a moderate effect size of d=.5. The power analysis can be found in the accompanying scripts.

- *For 1.2.3.: Sample size calculation is missing completely - for Linear Mixed Models you need specific sample sizes in order to make valuable statements, especially when interpreting random effects. Please complete.*

We completely agree that a proper power analysis should be reported in case of Linear Mixed Models. However, we reformulated our experimental design following reviewers' suggestions and finally decided to not fit models predicting liking from perceptual features in the context of the present RR.

In general, I think 1.2.3. can benefit from more detail: How are you going to decide which effects will remain in the final model? Step-wise LMM? Random slopes or interecepts? And so on...

You are right, our analyses plan for the linear mixed models was lacking details. Part of them would have been clearer with our (unfortunately missing!) analysis code.

6. *Major: About the MMI: As you are testing participants in two sessions anyways, I think the Beholder Index (as described in Hönekopp, 2006, https://doi.org/10.1037/0096-1523.32.2.199 and in Specker et al., 2020, https://doi.org/10.1371/journal.pone.0232083) is the more appropriate analysis method to account for the measurement error! From Specker et al. (2020): "The bi method estimates variance components that can be interpreted as the observed variance attributed either to the participant or the stimulus (see [20], p.2 for a comprehensive explanation of the estimation of the variance components). In order to do this, participants need to rate each stimulus twice. The repeated measure allows for estimation not just of how much participants agree with each other on a rating (shared evaluation) but also how much participants agree with themselves on the repeated rating (private evaluation)."*

Thank you very much for this valuable suggestion. We have included variance component analysis and computation of the beholder index in our analysis plan. They are listed as "supporting analyses" (Sections 1.3.3 and 2.4.3), since it is not straightforward to statistically compare beholder indices across categories, but we are certain these analyses will provide valuable insight into voice preferences.

7. *Minor: Is singing to infants only used cross-culturally or are its features also comparable across cultures? This is for me not clear from the intro.*

Sorry for the lack of clarity. In our revised manuscript this point is not so emphasized, but we would like to clarify this point here. The literature indicates both are true, that is, that lullabies are cross-culturally recognized (Mehr et al., 2018, 2019; Trehub et al., 1993; Yurdum et al., 2023), and that there are acoustic regularities depending on a song's function across cultures, so that the authors argue for universal form-function associations. This is also supported by findings describing cross-cultural acoustic regularities in infant-directed vocalizations (Cox et al., 2022; Hilton et al., 2022).

8. *Minor: When you discuss voice attractiveness, you could also mention studies which don't find effects of perceptual/acoustic features on voice attractiveness (or e.g., Mook & Mitchel, 2019, 10.1037/ebs0000128, who did find that averageness lowered voice attractiveness)*

Thank you for bringing this paper to our attention. We have included it in the relevant part of the introduction (line 61).

9. *Minor: Could you give some more information on which singing styles were employed in the Valentova et al. (2019) study?*

In the Valentova et al. (2019) study, participants sang Happy Birthday, spoke a short scripted salutation text, and both sang and spoke the text of excerpts of their (Brazilian and Czech) national anthems. We have included a summary of this information in the manuscript.

10. *Minor: Maybe focus more on individual vs. private taste in the introduction; you could give examples from face research or other aesthetic research (artworks etc.). Also, when you describe your previous studies, give more detail on how much was explained by individual differences. In the last paragraph before study aims and hypotheses (p. 4/5) make clear how these results apply to the different song & speech styles you are using.*

Thank you for these helpful suggestions. We have made the approach of quantifying share vs private taste clearer. We mentioned the (very small) amount of shared taste in our previous research with pop singing (we found highly idiosyncratic taste, as indicated by low interrater agreement of Krippendorff's alpha = .16 for the liking scale).
Also, we have more thoroughly covered research on the visual domain, and we have made the parallel to our styles of vocalizations clearer.

11. *Minor: Although you plan to do exploratory analyses in 1.1.3., I think you could be more precise. Which perceptual features are taken into account and why? What could you expect for e.g. the speaking voices based on previous results?*

You are right, that part was lacking precision. Since we reformulated our experimental design and removed exploratory analyses from the manuscript, this does not apply anymore.

12. *Minor: Can you maybe test hearing impairments instead of using self-report?*

Hearing ability is indeed very important here. We first thought about using an audiometer to ensure that participants had no hearing impairment, as has been done in several previous studies by our team, but in our experience, deficits were very rare among the recruited participants (since it is stated in the recruitment announcement that we are looking for participants without hearing impairment) and the few participants who were slightly below threshold were also self-reporting difficulties. Therefore, in order to keep the session as short as possible, we decided to use self-report only.

13. Minor: What biographical data do you collect and why?

We collect basic demographics such as age, gender, languages spoken, but also sexual orientation because of accounts of differences in voice attractiveness depending on gender/sex orientation. We plan to use this information for exploratory analyses.
That is now precisely described in the revised manuscript.

14. Minor: The three questionnaires could be explained in more detail why they are included, what they measure precisely (e.g. the Music Sophistication subscale) - maybe you could even mention sth about them in the introduction!

The reviewer is right that more details should have been provided. We now describe better the content of the Gold-MSI and additional questions. Note that we decided to remove the other two questionnaires and limit ourselves to the information we actually use, that is, only questions that connect directly to the study, such as particular preferences for opera and pop music, languages spoken and the mentioned demographics, and some additional information about participants' experience while doing the task.

15. Minor: Could you add some comparison within style (especially pop, which has low accuracy?) instead of just between styles?

We are not sure what you mean. We did compare the acoustics of different singing styles in a separate study (Bruder & Larrouy-Maestri, 2023) to examine singer's versatility. In this context, we investigated the accuracy of style recognition by participants in the validation experiment (also briefly described in the current manuscript) in relation to singers' classical training. We observed that the pop style had lower accuracy of recognition than the other styles, and that pop and lullaby performances were sometimes mistaken for each other. The lower proportion of correct recognition of pop performances might be related to the broad and loosely defined meaning of the term "pop", as well as to singers' lower proficiency/experience in this style - though note that the proportion of correct style recognition was way above chance-level performance for all styles. We considered removing the items with lowest recognition from the stimulus set planned for the current study (for instance, removing all items with proportion of recognition under a certain threshold), but another problem would then appear. Since the items with lowest recognition come from different singers, in different melodies, and in different styles, removing them would leave “holes” in the stimulus set - we would lose the very appealing feature of the set being fully matched, that is, that all voices perform the same material in all styles.

In any case, we plan to run exploratory analyses about the relationship between how “recognizable” (as a proxy to how typical) a stimulus was in the validation experiment and how much it was liked in the currently proposed experiment.

16. *Minor: Why are the questionnaires presented half-way through the experiment? Maybe you could present them at both sessions to replicate?*

In our last study (Experiment 2 of Bruder et al, 2023), we presented questionnaires half-way through the experiment to let participants "rest" from the repetitive task of rating 96 stimuli in terms of liking as well as in other 10 perceptual scales (in only one testing session). We were planning to apply the same logic here. However, in our revised experimental design (only including liking ratings), we propose to collect questionnaires in the end of the first session for simplicity.

Based on available literature, we would expect participants’ answers to our proposed questionnaires to be consistent across the two testing sessions. Müllensiefen et al. (2014) reported very high test-retest reliability of their General Music Sophistication subscale (Pearson $r_{test\text{-}retest}$ = .97) for participants tested between 10 and 64 days apart (average interval of 23 days, SD = 9.2). The same seems to apply for preferences for particular music genres (Lawendowski, 2013- conference paper). This suggests that collecting these data two times might not be necessary.

17. *Minor: How is the randomization conducted?*

In the revised experimental design, each style of vocalization will be presented in a separate block of trials (in counterbalanced order across participants). Within each block, performances of three different melodies by 22 singers (for a total of 66 trials per block) will be presented in random order. We made that clear in the revised manuscript.

18. *Minor: I feel like it could be advantageous to always let participate rate liking first before rating all other dimensions to get a more spontaneous rating there?*

We fully agree that the different perceptual scales may increase participants' awareness of certain aspects of the performances and thus influence their liking behavior. However, as discussed in response to previous comments, we decided to focus only on liking ratings for now, which means that this is no longer an issue. But we would like to note that we are currently preparing a specific experiment to test this and would be happy to share the results when the study is completed.

Formalities:

1. *There are a few grammar issues (especially regarding prepositions)*
2. *Report is not APA-conform; the reference list definitely needs a review! Also: Headings, Table descriptions, Figure descriptions etc.*

Thank you for pointing that out. We have thoroughly revised those issues.

References

Albouy, P., Mehr, S. A., Hoyer, R. S., Ginzburg, J., & Zatorre, R. J. (2023). *Spectro-temporal acoustical markers differentiate speech from song across cultures* [Preprint]. Neuroscience. https://doi.org/10.1101/2023.01.29.526133

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Brooks/Cole.

Babel, M., McGuire, G., & King, J. (2014). Towards a more nuanced view of vocal attractiveness. *PLoS ONE*, *9*(2), e88616. https://doi.org/10.1371/journal.pone.0088616

Brown, S. (2000). The 'Musilanguage' model of language evolution. In *The Origins of Music* (eds S. Brown, B. Merker, and N. L. Wallin, pp. 271–300). MIT Press.

Bruder, C., & Larrouy-Maestri, P. (2023). Classical singers are also proficient in non-classical singing. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1215370

Bruder, C., Poeppel, D., & Larrouy-Maestri, P. (2023). *Perceptual (but not acoustic) features predict singing voice preferences* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/qvp8t

Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G., & Fusaroli, R. (2022). A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behaviour*, *7*(1), 114–133. https://doi.org/10.1038/s41562-022-01452-1

Fischinger, T., Kaufmann, M., & Schlotz, W. (2020). If it's Mozart, it must be good? The influence of textual information and age on musical appreciation. *Psychology of Music*, *48*(4), 579–597. https://doi.org/10.1177/0305735618812216

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., Simson, J., Knox, D., Glowacki, L., Alemu, E., Galbarczyk, A., Jasienska, G., Ross, C. T., Neff, M. B., Martin, A., Cirelli, L. K., Trehub, S. E., Song, J., Kim, M., … Mehr, S. A. (2022). Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour*, *6*(11), 1545–1556. https://doi.org/10.1038/s41562-022-01410-x

Hird, E., & North, A. (2021). The relationship between uses of music, musical taste, age, and life goals. *Psychology of Music*, *49*(4), 872–889. https://doi.org/10.1177/0305735620915247

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology:*

Human Perception and Performance, *32*(2), 199–209. https://doi.org/10.1037/0096-1523.32.2.199

Lawendowski, R. (2013, June 11). Temporal stability of music preferences as an indicator of their underlying conditionings. *Proceedings of the 3rd International Conference on Music & Emotion*. ICME3, Jyväskylä, Finland.

Leongómez, J. D., Havlíček, J., & Roberts, S. C. (2022). Musicality in human vocal communication: An evolutionary perspective. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1841), 20200391. https://doi.org/10.1098/rstb.2020.0391

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. V., Simson, J., Bainbridge, C. M., Pinker, S., O'Donnell, T. J., Krasnow, M. M., & Glowacki, L. (2019). Universality and diversity in human song. *Science*, *366*(6468), eaax0868. https://doi.org/10.1126/science.aax0868

Mehr, S. A., Singh, M., York, H., Glowacki, L., & Krasnow, M. M. (2018). Form and function in human song. *Current Biology*, *28*(3), 356-368.e5. https://doi.org/10.1016/j.cub.2017.12.042

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

Ozaki, Y., Tierney, A., Pfordresher, P., Mcbride, J., Benetos, E., Proutskova, P., Chiba, G., Liu, F., Jacoby, N., Purdy, S., Opondo, P., Fitch, T., Hegde, S., Rocamora, M., Thorne, R., Nweke, F. E., Sadaphal, D., Sadaphal, P., Hadavi, S., … Savage, P. E. (2022). *Globally, songs and instrumental melodies are slower, higher, and use more stable pitches than speech [Stage 2 Registered Report]* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/jr9x7

Phillips, E. (2023). *A review of the speech-music continuum and its categorization: Evolution, form, and function* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/b2tkf

Rhodes, G. (2006). The Evolutionary Psychology of Facial Beauty. *Annual Review of Psychology*, *57*(1), 199–226. https://doi.org/10.1146/annurev.psych.57.102904.190208

Sharma, B., Gao, X., Vijayan, K., Tian, X., & Li, H. (2021). NHSS: A speech and singing parallel database. *Speech Communication*, *133*, 9–22. https://doi.org/10.1016/j.specom.2021.07.002

Trehub, S. E., Unyk, A. M., & Trainor, L. J. (1993). Adults identify infant-directed music across cultures. *Infant Behavior and Development*, *16*(2), 193–211. https://doi.org/10.1016/0163-6383(93)80017-3

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

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

Vessel, E. A., & Rubin, N. (2010). Beauty and the beholder: Highly individual taste for abstract, but not real-world images. *Journal of Vision*, *10*(2), 1–14. https://doi.org/10.1167/10.2.18

Vessel, E. A., Stahl, J., Maurer, N., Denker, A., & Starr, G. G. (2014). *Personalized visual aesthetics* (B. E. Rogowitz, T. N. Pappas, & H. De Ridder, Eds.; p. 90140S). https://doi.org/10.1117/12.2043126

Yurdum, L., Singh, M., Glowacki, L., Vardy, T., Atkinson, Q. D., Hilton, C. B., Sauter, D., Krasnow, M. M., & Mehr, S. A. (2023). Universal interpretations of vocal music. *Proceedings of the National Academy of Sciences*, *120*(37), e2218593120. https://doi.org/10.1073/pnas.2218593120