Savage et al. response letter to Logan PCI-RR triage

Thanks so much for this extremely quick and detailed set of suggestions. Due to the large number of coauthors and short time frame we will need to wait to address many of the more substantive suggestions during the next revision round, but we have tried to address as many as we could quickly do now. In particular, we have tried to clarify that the pilot data analysed DO come from the new Savage et al. 2025 experiment design (they are the same pilot data as the pilot data analysed by Savage et al. in their Fig. 4).

- Abstract: needs to discuss the role this study plays in the larger literature, and a statement at the end about what the results will do to drive the field forward. The Savage et al (2025) abstract is a good example of a complete abstract.

Added to abstract:

> **Regardless of results, our multi-site large-scale replications will enhance our understanding of cross-cultural relationships between music and language and provide a template for equitable collaboration using PCI-RR's Programmatic track.**

- Introduction: needs more background on the study topic, more discussion of the hypotheses, what their implications are, and what your interpretations will be given all possible outcomes (positive association, negative association, or no association), and how the results will advance knowledge in this field. This Stage 1 will be the basis for many Stage 2s so it is important to have a very solid and thorough introduction for everyone to work from when writing their Stage 2. It should read like a regular and complete introduction, plus the additional details about the big team science (in perhaps a subsection of the introduction?). The Savage et al (2025) introduction is a good example for how to make the current intro complete.

(Will address in revision - note that key details are currently summarised very briefly in Table 1, but we will expand on these more fully in our revision)

- Lines 62 and 145: The analyses require two recordings per subject. How will you isolate individual singing in recordings if these individuals are singing and speaking in a group?

Added this clarification:

> **Audio recordings from all conditions will be made publicly available for analysis/replication. Participants will also be monitored by video. Video will not be published, but will be used by each local team to ensure experiment instruction compliance and to match singing/speaking recordings for individuals within the group.**

- Lines 70-73: how and why did the automated method for analyzing the recordings differ from the experimenter's analysis of the recordings? It is an interesting point that the automated version came to the opposite conclusion from the experimenter's version and it needs more detail. It is well described in Ozaki et al, so please flesh out this description here as well.

(Will address in revision)

- Methods: need to give more details about the protocols that you are using from the other studies so this Stage 1 can stand alone without readers needing to refer to 2 other publications to understand. It will also make it easier for authors to write the Stage 2s. For example, how did experimenters choose songs, speech text, participants, etc.?

(Will address in revision)

- Lines 117-120: What program(s) will the audio data be transcribed into? What is the protocol for segmenting the recordings into acoustic units? What is the definition of an acoustic unit? What program(s) will be used to replicate Ozaki et al's analyses?

We have updated the Appendix containing the link to a video of the segmentation protocol with additional details as follows (we can also write out some of these key details during the next revision):

> APPENDIX S1
>
> Segmentation protocol/tutorial **[this is the version used by Ozaki et al. using Sonic Visualiser - we plan to create a new version specifically for this new protocol updating Ozaki et al.'s video to use Praat]:**
> https://drive.google.com/file/d/1YOiobvoxaM4txdAJDVeLjc--oNLiBb5n/view

- Line 144: “15-30 individuals per site singing a pre-chosen song from their language/culture (in unison in a group and monophonically alternating line by line with other participants)”. Does this mean that every person will have sung each line solo for the recording? If so, then it seems like these solo recordings are what you would use in the analyses because it would be for each individual participant and audio from other participants speaking/singing at the same time would not also be included in the recording.

- Lines 141-147: Is the “recitation condition” the speaking recording against which the singing recording will be compared? It seems like it is a tighter control if the speaking recordings are of the participants speaking the lyrics of the song they sing. That way there aren’t differences in the variables of interest due to different words being used. This isn’t my area of expertise though so perhaps there are reasons not to do this. It looks like this is much better described in Ozaki (page 4), so please write an analogous description here. There are a variety of different recordings of both singing and speaking, but only one version of each will be analyzed (lines 149-151). Please clarify this section to indicate why the additional song and speech recordings are being recorded and what they will be used for (I see that this is well described in Ozaki, so please describe here. Also, if you are only analyzing 2 of the types of recordings in this registered report, then perhaps omit the other recording types for clarity?). How will you identify individuals in the group recordings?

We realise this is a bit confusing, sorry. Will try to rewrite in revision in a way that tries to explain more completely without becoming even more confusing. The important thing here is that only the alternating singing condition and free conversation conditions provide one person singing/speaking at a time to allow us to isolate and directly compare within-participant singing/speaking audio.

- Line 154: please clarify what “outcome-neutral criteria” means with regard to this study for readers who are not familiar with the term. This will help the reader better understand when you say “Thus it is possible that some data collected for that study will pass those outcome-neutral controls, but fail to provide reliable audio data”, which is unclear at the moment.

- Line 160: “those audio recordings will be subject to a separate set of outcome-neutral inclusion criteria”. Please describe what these criteria are.

- Line 168: what are the “minimum standards of quality”? It will be important to have thresholds with numerical values assigned so it is clear what is inside and what is outside the minimum standards of quality. This will be useful to have in the Stage 1 for when the experimenters are implementing the protocol to collect and analyze the data.

(Will address in revision)

- Line 207: typo “analyses” should be “analyze”

Fixed, thanks

- Line 238: Do you plan to conduct interrater reliability (IRR) in this study as well? It seems like it would be beneficial to train the coders to a certain level to ensure a minimum IRR before they code the data involved in the current study (for both singing and speaking, and for the language(s) they are going to code). This seems particularly needed for the current study because the audio recordings will come from a group setting, rather than a single individual. It would be useful to describe the data collection and data analysis protocols in great detail in the Stage 1 so that all teams carry out the same steps in the same way. For an example of how I do this in my lab (including R code), see Supplementary Material 3 in Logan et al. (2023).
Logan, Corina; McCune, Kelsey; LeGrande-Rolls, Christa; Marfori, Zara; Hubbard, Josephine; Lukas, Dieter. Implementing a rapid geographic range expansion - the role of behavior changes. Peer Community Journal, Volume 3 (2023), article no. e85. doi : 10.24072/pcjournal.320. https://peercommunityjournal.org/articles/10.24072/pcjournal.320/

Will discuss with group and consider adding in revision (will need to consider feasibility issues). We will definitely train coders using the tutorial video (Appendix S1).

- Statistical analysis: please include a description of the models you will use to analyze the hypotheses. The Savage et al (2025) analysis section is a good model to follow.

(Will address in revision)

- Line 279: “Figure 3 shows our proposed analyses (replicating Figs. 4-5 from Ozaki et al.) for the same pilot experiment shown in Savage et al.’s Figure 41”. It isn’t clear what the pilot experiment in Savage et al. (2025) was, or how you are replicating the figures from Ozaki (e.g., with what data). Please explain where these data come (I think Ozaki?) and describe more about how this is a simulation using data from Ozaki to determine minimum sample sizes for the current study (if I understood this correctly).

- Lines 282-284: “This suggests that the various changes made from Ozaki et al.’s design (e.g., a single microphone recording multiple individuals engaged in alternating singing/free conversation) did not compromise our ability to accurately replicate their analyses”. However, I thought that the data in this pilot were from Ozaki (recordings from individuals), in which case, you wouldn’t be able to say how the changes you made to the protocol (i.e., group recordings) will change the detection of the variables of interest. The same comment applies to the next sentence in the paragraph as well: “The general trends (i.e., song again appearing higher, slower, and more stable than speech) also indicate that any changes in experimental design are unlikely to result in major changes to results.”

- Lines 291-293: the figure 3 legend makes it seem like the pilot data actually come from Savage et al (2025), however that isn’t possible because that data have not been collected yet (as indicated in the Report Survey at PCI RR where the data collection start date is in April 2025 and the level of bias is level 6, which means no data have been collected yet). A clarification throughout the methods and particularly in this pilot section about where the data come from and what you are comparing it to will be useful.

There appears to be important confusion here - these pilot data DO come from Savage et al. 2025, but only from their PILOT data (i.e., not data that will be produced in their Stage 2 confirmatory analyses). To make sure this is clear we have added the bold text to the following:

> Figure 3 shows our proposed analyses (replicating Figs. 4-5 from Ozaki et al.) for the same pilot experiment shown in Savage et al.’s Figure 4[1] (i.e., n=14 participants from Auckland, New Zealand singing/speaking in English **in three groups of 4-5 people at a time in June/July 2024). Note that these pilot experiments will not be used for confirmatory analyses in Stage 2 reports resulting from either ref. 1 or the current protocol - they are only used for pilot analyses.**

- Lines 294-297: it is confusing which effect sizes go with which study. Perhaps a table is more useful here, with a column for Savage et al 2025 and a column for Ozaki et al?

Added:

Table 2. Comparison of effect sizes (translated Cohen’s *D*) for the three proposed features from Ozaki et al.’s original study (n=75 participants[2]) with pilot data from the current protocol (n=14 participants from Savage et al.[1])

Feature	Original effect size	Pilot effect size
Pitch height	1.6	0.8
Temporal rate	1.6	1.8
Pitch stability	0.7	0.4

- Line 300: what does “single shared tonal centre” mean? And why were all singers forced to use only one?

Edited to read:

> **…singing in a group context (even when each singer sings one solo line at a time) usually results in all singers matching their singing to a single shared key (tonal centre)...**

- Data at https://github.com/comp-music-lab/manyvoices3/tree/main: the README file (or some other metadata file that you provide) needs to have detailed information about all of the files at GitHub, what they are used for, and what they correspond to in the Stage 1. For example, there are many csv files in the folder pitch (https://github.com/comp-music-lab/manyvoices3/tree/main/data/pitch). There seems to be a file naming convention, but it is not clear what this is, so it needs to be explained. Also, a list of software that can run the various analysis pieces would be useful. I tried to run some of the .m files in R, but it didn’t work (e.g., https://github.com/comp-music-lab/manyvoices3/blob/main/simulation_analysis/cwtdiff.m).

(Will address in revision - or earlier if possible! But this won't affect the manuscript so please go ahead and release the updated manuscript to reviewers and we'll update GitHub as soon as we can.)

I see that appendix S1 is referenced in the Stage 1 pdf, but I don't see the appendix as part of the pdf. Can you please append it to the pdf and update it at PsyRxiv? Unless it refers to the appendix in Savage et al. (2025)? In which case, the text in the current stage 1 should be updated and a link to the appendix should be given.

Appendix S1 is simply a short link to the segmentation tutorial video (previously Lines 389-392, now lines 405-410). We have now updated it to give more context as discussed above.