School of Psychology
Science Centre Level 2, Building 302,
23 Symonds Street Auckland, New Zealand
(+64) 21-672-013

2 April, 2025

RE: *A Programmatic Stage 1 Registered Report of global song-speech relationships replicating and extending Ozaki et al. (2024) and Savage et al. (2025)*
(*https://doi.org/10.31234/osf.io/c2dba*)

Dear Dr. Logan,

We appreciate your invitation to revise and resubmit our Stage 1 Registered Report protocol based on the constructive comments of yourself and the three Reviewers. We are grateful for the chance to use their feedback to modify our planned protocol to enable us to reach stronger conclusions following our eventual Stage 2 data collection and analysis.

Most changes simply required additional clarification (including two new diagrams to clarify the experimental and analysis plans). There are two substantive changes to flag:

1) **We have added plans to measure inter-rater reliability of annotations**
2) **We have changed the order of alternating vs. group singing conditions in one experiment to address Dr. Sadakata's point that this order could potentially affect results**

Changing this second point required a minor change to Savage et al.'s In Principle Accepted experimental protocol (which we have done, notifying that Recommender, Dr Katherine Moore). Since we had not yet begun Stage 2 data collection for that experiment (as clarified below), this will not affect any of our plans.

We have appended a version with tracked changes to this response letter for your convenience.

We feel that the review process has substantially improved our Stage 1 protocol. We hope you will find the revised protocol ready for In-Principle Acceptance and Stage 2 data collection.

Sincerely,

Patrick Savage (on behalf of the authors)

Full reviews/decision:

Recommender's decision/summary (Corina Logan):

Decision for round #1 : *Revision needed*

Revise

Dear Dr Savage and co-authors,

Thank you for your submission to PCI RR. I appreciate that you are using the Programmatic Registered Report innovation to its fullest potential and I love that you are using it to improve the equitable sharing of co-authorship. I'm glad to be involved in this process!

We are delighted you share our excitement about this central part of our proposal.

I have received feedback from three reviewers and, combined with my own feedback, I welcome a revised version of your Stage 1. See below for the reviewer feedback. My main comment is that the abstract, introduction, and methods need more detail to be replicable by team members and others (see review criterion 1D at https://rr.peercommunityin.org/help/guide_for_recommenders#h_67596462364016136433909 05).

We have now substantially expanded the details in the abstract, Table 1 (Registered Report Design Planner), introduction, and methods.

Additionally, I thank you for having a quick, pre-review back and forth with me, which resulted in a partial revision based on some of the changes I suggested. I include your author response to my comments in the PDF below, which shows my detailed comments on your submission and how you have already addressed some of them.

For simplicity, we have included below only your comments we did not previously fully address, along with our new responses to them.

I note that the question about the level of bias for this submission (currently proposed as a level 6) is still unanswered. From Savage et al (2025), it looks like your planned data collection start date was 1 Dec 2024, which means that some of the data (the recordings) for the current submission is already being collected. That means that this submission would be a level 4 or below. Please explain what stage the data collection is at and how visible the data are to the authors.

Our original goal of Stage 2 data collection beginning around 1 Dec 2024 was based on a rough estimate of when we thought we might achieve In Principle Acceptance. However, the actual date of In Principle Acceptance was not until the end of January 2025, with Level 6 bias control (https://rr.peercommunityin.org/user/recommendations?articleId=890). We had not yet begun any Stage 2 data collection for Savage et al. (2025) when we submitted this Programmatic Stage 1 protocol to PCI-RR on 14 February 2025. Our first Stage 2 data collection for that project began on 21 March 2025, for Mandarin speakers in London not

involved in this Programmatic protocol). However, all of the 26 sites that are part of the current Programmatic Registered Report (see Table 2) have committed to continue waiting to collect data until this also receives In Principle Acceptance in order to maintain Level 6 bias control.

I look forward to receiving your revision.

All my best,

Corina

by ***Corina Logan****, 04 Mar 2025 12:58*
Manuscript: ***https://doi.org/10.31234/osf.io/c2dba_v3***
version: 1

Logan previous comments addressed in new submission:

- 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.

We have substantially expanded the Introduction as requested (copied below with new sections highlighted in bold):

> INTRODUCTION
>
> Music and language are two human cultural universals found in all known societies: separately (e.g., instrumental music, speech), but also together in the form of songs with words[3–10]. Previous research in fields including musicology, linguistics, psychology, anthropology, and neuroscience has identified neural, acoustic, and behavioural relationships between song and speech[11,8,12,13,9,14]. However, most previous research has been limited to speakers of English and other European languages, limiting the generality of conclusions that can be drawn[8,15,16].
>
> **A key unresolved question is what, if anything, consistently distinguishes song from speech across languages? Steven Pinker famously dismissed music as an evolutionarily "useless" byproduct of adaptative traits such as language[17], while others have argued that the regular pitches and rhythms of music facilitate adaptative functions such as bonding individuals together or signaling group membership beyond the capacities of language[18,19]. However, these debates have mostly been conducted in the absence of direct cross-cultural comparisons of actual singing and speaking[20–23].**

Recently, Ozaki et al. compared audio recordings of singing and speaking from their 75 coauthors speaking 55 languages, concluding that "*Globally, songs and instrumental melodies are slower and higher and use more stable pitches than speech*" **and speculating that "*the slower and more stable pitches may facilitate synchronization, harmonization, and ultimately bonding between multiple individuals*"**[2]. However, their coauthors were mostly researchers and professional musicians who were not representative of general speakers/singers of their languages, so the degree to which their findings would generalise to other speakers of their languages remains unclear[24]. While this limitation was mitigated to some extent by comparison of singing and speaking recordings from separate cross-linguistic databases[2,25–27], these databases did not include annotated segmentations into acoustic units (e.g., syllables/notes), meaning it was not possible to directly replicate or compare with Ozaki et al.'s analyses. **And because Ozaki et al. only included one or a few speakers of each language and averaged their results across many languages, it is possible that some of their results may display different effects in different languages. For example, tonal languages such as Mandarin or Yoruba could conceivably use more stable spoken pitches, while "mora-timed" languages such as Japanese could be faster[28].**

Another key limitation of previous datasets is that they included only solo singing/speaking, whereas most singing and speaking throughout the world tends to be done in groups[10,29,30]. To overcome this, we have designed a new study in collaboration with over 80 researchers aiming to collect data on group singing and speaking in diverse languages from over 1,500 participants across over 50 different sites around the world[1]. However, there remains the challenge of annotating all this data in an efficient and equitable way.

Equitable coauthorship in global collaboration:

One factor underlying the annotation issue is the broader challenge in big team science of ensuring equitable credit and authorship for all collaborators at all locations, rather than only having them listed as middle authors in a large coauthored publication (or not listed as coauthors at all)[31–35]. High-quality segmentation of cross-cultural audio recording corpuses requires many different researchers who are speakers of diverse languages to spend substantial time manually annotating audio recordings[2]. While one might hope that automated segmentation technology might reduce or eliminate this barrier, Ozaki et al.'s analyses found that using automated segmentation tools are not yet reliable, and in fact would have led to incorrect conclusions:

> **While automatic segmentation can be effective for segmenting some musical instruments and animal songs [e.g., percussion instruments and bird song notes separated by microbreaths], we found that they did not provide satisfactory segmentation results compared to human manual annotation for the required task of segmenting continuous song/speech into discrete acoustic units such as notes or syllables…. For example,**

> **Mertens' automated segmentation algorithm used by Hilton et al. mis-segmented two of the first three words "by a lonely" from the English song used in our pilot analyses ("The Fields of Athenry"), oversegmenting "by" into "b-y," and undersegmenting "lonely" by failing to divide it into "lone-ly"...***if we had used this automated method, then we would have mistakenly concluded that there is no meaningful difference in IOI [Inter-Onset Interval] rates of singing and speech…****collaboration with native/heritage speakers who recorded and annotated their own speaking/singing relying on their own Indigenous/local knowledge of their language and culture allowed us to achieve annotations faithful to their perception of vocal/instrumental sound production that we could not have achieved using automated algorithms…This highlights that equitable collaboration is not only an issue of social justice but also an issue of scientific quality***"[2]).

Using *Peer Community In Registered Reports*' Programmatic model[36,37], we aim to overcome these challenges by incentivizing each local team to segment and analyse data from their own language/culture by providing them with the opportunity to publish a first-authored article. We propose to create up to 27 Stage 2 Registered Reports (Table 1) that all follow the basic protocol of this Stage 1 Registered Report. These teams represent a subset of the 60 global teams that have agreed to collect singing/speaking data from 15-30 participants each as part of a broader study on the behavioural effects of singing/speaking on social bonding[1].

By unifying these Stage 2 Registered Reports around a small shared set of three hypotheses for confirmatory testing, this should allow for coherence across different teams using shared methods, while also giving each team the flexibility to add additional exploratory analyses according to their own interests. For example, some sites are based in ethnomusicology departments and may add qualitative ethnographic analyses; others are based in psychology departments and may add extra analyses of demographic data; others are based in computer science departments and may add extra acoustic analyses. However, all teams will collect, analyse and report the same basic confirmatory hypothesis testing replicating Ozaki et al.'s original acoustic comparison of song and speech[2].

Hypotheses. We hypothesize that Ozaki et al. 's findings of key differences between singing and speaking will replicate in all languages and all sites tested. Specifically:

1) Singing uses higher pitch than speech
2) Singing is slower than speech
3) Singing uses more stable pitches than speech

For each site/Stage 2 report, we will conclude whether or not each of Ozaki et al.'s three key findings (regarding tempo, pitch height, and pitch stability) generalise to their given language/location. We will also include a meta-analysis comparing all sites with Ozaki et al.'s original results to conclude whether their findings generalise across all studied

languages/locations. **For each of the three features in each language, we will conclude whether songs are significantly higher/faster/more stable than speech (replicating Ozaki et al.), significantly equivalent (contradicting Ozaki et al.), or inconclusive (if neither null hypothesis testing nor equivalence testing are statistically significant; see Table 1).**

Since this is a Programmatic Registered Report where one Stage 1 protocol will result in multiple Stage 2 outputs, it is possible that different Stage 2 outputs will produce different results for different languages. This will allow us to evaluate criticisms that global analyses of cross-cultural trends fail to address the importance of internal diversity ("*How many exceptions are researchers willing to ignore?*"[28]).

Except for the Stage 2 output combining all studies (#27 in Table 1), each Stage 2 will focus its confirmatory analyses on the results of its own analysis of its own focus language. #27 will replicate Ozaki et al.'s cross-linguistic meta-analysis approach to analyse average trends across all languages, which can be compared with the results of each individual Stage 2 reports #1-26 to achieve a much broader evaluation of the cross-linguistic replicability and generalisability of Ozaki et al.'s original results. Comparison of specific differences between languages will be reserved for exploratory analysis (since statistical power for such comparisons will be limited by the relatively small sample size of n=15-30 participants per language).

To ensure maximal consistency across Stage 2 reports, all Stage 2 reports will restrict their confirmatory analyses and statistical hypothesis testing to only these three hypotheses. They are welcome and encouraged to explore additional analyses, but must ensure these conform to PCI-RR's Stage 2 criterion 2D[38]:

> ***2D. Where applicable, whether any unregistered exploratory analyses are justified, methodologically sound, and informative***

- 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.

We have substantially expanded our quote from Ozaki et al., adding the following bolded sections:

> ***While automatic segmentation can be effective for segmenting some musical instruments and animal songs [e.g., percussion instruments and bird song notes separated by microbreaths], we found that they did not provide satisfactory segmentation results compared to human manual annotation for the required task***

of segmenting continuous song/speech into discrete acoustic units such as notes or syllables…. For example, Mertens' automated segmentation algorithm used by Hilton et al. mis-segmented two of the first three words "by a lonely" from the English song used in our pilot analyses ("The Fields of Athenry"), oversegmenting "by" into "b-y," and undersegmenting "lonely" by failing to divide it into "lone-ly"...*if we had used this automated method, then we would have mistakenly concluded that there is no meaningful difference in IOI [Inter-Onset Interval] rates of singing and speech…* ***collaboration with native/heritage speakers who recorded and annotated their own speaking/singing relying on their own Indigenous/local knowledge of their language and culture allowed us to achieve annotations faithful to their perception of vocal/instrumental sound production that we could not have achieved using automated algorithms…This highlights that equitable collaboration is not only an issue of social justice but also an issue of scientific quality*****).**

- 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.?

We have now added these details the following new "Song/speech selection and participant inclusion criteria" section and in response to reviewer questions below. In general we have tried to exactly quote relevant parts of previous protocols where possible to ensure consistency:

Song/speech selection and participant inclusion criteria:

Each site in Savage et al.[1] will recruit 15-30 participants and choose its own song (cf. Table S1 from ref. [1]) and conversation prompt using the following criteria. Note that the need to recruit participants to sing together in groups means it is not feasible to allow each participant to choose their own song as Ozaki et al. did:

Participant inclusion criteria:
Each site will recruit participants who meet the following inclusion criteria:
-Age 18 or over
-Able to sing the song chosen for that site (with lyrics provided)
-Able to converse in the same language its lyrics are written in
-Have access to a phone or other device that can scan QR codes
-Willing to have their singing/speaking voice recorded and shared publicly (without being identified by name)

Song selection criteria

Each site has chosen a song that would be appropriate for their language/culture. The criteria for choosing a song were:

-lyrics are mostly in the same language that participants will use for their group

conversation (some lyrics in other languages or meaningless vocables like "la la" are acceptable, but should not make up the majority of the song)

-should be easy for most potential participants from that society to sing together in synchrony (e.g., unison, homophony) with karaoke-style pre-recorded instrumental accompaniment without needing to practise ahead of time. If possible, this should be in the form of a karaoke-style video with plain background and lyrics that appear in real-time to help the participants to sing at the right time, with no guide melody (e.g., https://youtu.be/OhRUYf_yn_s?si=eL4mt_-utRwqrFMj&t=10). If pre-recorded instrumental accompaniment would not be appropriate for a given site/society, an a cappella (unaccompanied) song may be chosen instead.

-should be the kind of song that would be appropriate to sing by young adults who don't already know each other as a short "ice-breaker" exercise. As such, songs that might easily become awkward, embarrassing, or offensive should be avoided (e.g., children's songs, songs with polarising content or associations such as national anthems or religious songs). However, these factors may vary from site to site (e.g., for some communities a national anthem or religious song might be the best choice, while in others it might be the worst). The experimenters from each site should interpret this on the basis of their own local knowledge.

-the song should take between 2-3 minutes to sing (you are welcome to modify the number of verses/choruses (including repeating the song) to make this happen

-if the song has instrumental interludes/introductions/outros, these should not be longer than 1 minute total and there should still be 2-3 minutes of singing time not including these instrumental sections.

Conversation ice-breaker question criteria:

Each team will choose their own unique ice-breaker question for the conversation condition (this can be taken directly from one of the following lists, adapted from them, or newly created themselves, but teams should all choose different questions):
https://www.mural.co/blog/icebreaker-questions
https://museumhack.com/list-icebreakers-questions/
https://www.parabol.co/resources/icebreaker-questions/

Criteria for questions:
-Should not be about music/singing
-Should not use words/concepts that will be rated to create our dependent variable (i.e., "team", "similar", "trust", "close", "ties", "common").
-Should not ask sensitive/personally identifiable information (e.g., name, address, birthday, religion, sexuality, etc.)
-Should be capable of short answers (5-15 seconds per person)

- 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 added video tutorials for the segmentation protocol as follows:

APPENDIX S1: Video tutorials

A video tutorial showing how to use the free software Praat[47] to segment acoustic units (e.g., syllables/notes) from a pilot experiment recording containing multiple participants singing/conversing is available here: https://drive.google.com/file/d/1Nz4h-JSk1d3Z_NNiXN1UEpv3TVTBefdx/view?usp=sharing. The video used by Ozaki et al. showing how to align onsets based on perceptual centers ("P-centers") is available here: 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?).

Sorry for the confusion. We have now re-written this section and added a figure for clarity as follows:

Each of the three groups will engage in four conditions: group singing, alternating singing, conversation, lyric recitation. In contrast, Ozaki et al.'s confirmatory analyses compared 75 individuals singing a traditional song solo and then describing the meaning of the lyrics (also solo), comparing the first 20s of singing/speaking for each individual. (Ozaki et al. also included a solo lyric recitation condition and an instrumental melody condition, although these were not included in their confirmatory analyses).

In order to maximize comparability with Ozaki et al.'s analyses, we will focus our confirmatory analyses on only the conversation and alternating singing conditions (Fig. 2), as these are the ones with monophonic recordings of individual singing/speaking voices to

enable comparison of singing vs. speaking for each individual. (Any comparisons with the unison singing and lyric recitations conditions will be reserved for exploratory analyses.)

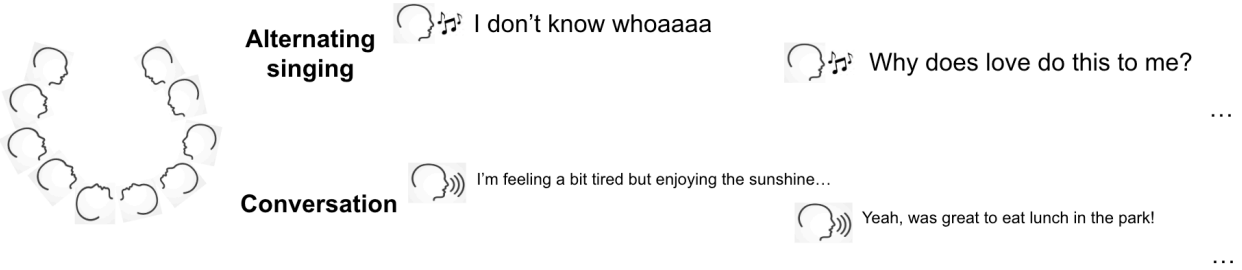

Figure 2. Schematic overview demonstrating an example of the two conditions analysed in confirmatory analyses. Here, only the first two participants are shown singing or speaking sequentially, but the total number of participants will be between 5-10 per experiment. Text columns #1 and #2 represent the first and second phrase of sequential singing/speaking. This example shows lyrics for "Why Does Love Do This To Me?", the song chosen for participants using New Zealand English, and hypothetical conversation based on the ice-breaker prompt "How is your week going?", but note that the actual song and conversation prompt will be different (and generally in a different language) at each site. (See Savage et al.[1]'s Fig. 1 for an illustration of the lyric recitation and synchronised singing conditions not included in the current confirmatory analyses.)

Note that, while Savage et al. compare social bonding effects of these different conditions using a between-participant design, our acoustic analysis proposed here instead compares singing vs. speaking for each individual participant in a *within*-participant design (i.e., comparing the same person's singing voice with their speaking voice, following Ozaki et al.'s original acoustic analyses).

…

At each site, the 15-30 participants will be randomly assigned into one of three groups. Each group completes the same four conditions (conversation, monophonic singing, unison singing, lyric recitation) but in different orders. When the (unaccompanied) monophonic singing condition follows the unison singing accompanied by karaoke-style accompaniment, participants may be influenced by having just heard and sung at the key and tempo matching this accompaniment. Likewise, it is possible that people may sing/speak differently depending on whether they have a conversation before or after singing. For these reasons, Savage et al. counter-balanced the order of conditions in the three participant groups as follows, enabling exploratory analyses of potential order effects:

-Group 1: 1) conversation, 2) monophonic (alternating) singing, 3) unison singing, 4) lyric recitation

-Group 2: 1) unison singing, 2) lyric recitation, 3) monophonic (alternating) singing, 4) conversation,

-Group 3: 1) lyric recitation, 2) unison singing, 3) conversation, 4) monophonic (alternating) singing

- 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.

We have expanded and re-written this section, as follows:

> **Outcome-neutral criteria ("*designed prior to knowledge of the results and …independent of the main study hypotheses*"[39]):**
>
> **Savage et al.'s experiment will employ the following outcome-neutral exclusion criteria[1]**
>
> ***-Participants who fail to show up on time at the agreed location***
>
> ***-Participants who fail to complete the experiment and submit the Qualtrics survey***
>
> ***-Participants who are unable to complete the singing/speaking task in the specified language***
>
> ***-Participants who fail the attention check***
>
> ***-Participants with any confirmatory dependent variable's data missing or corrupted due to technical glitches***
>
> ***-Participants with mean baseline social bonding scores of >80/100 (to avoid ceiling effects)***
>
> ***-Duplicate submissions by the same participant***
>
> ***-All participants from groups where "Instruction compliance" for the main experimental task (first condition) is judged unacceptable by the experimenter (<25 out of 100)***
>
> ***-Sites where useable data are only collected from fewer than 15 participants across all 3 groups***
>
> **Savage et al.'s criteria are focused on their confirmatory analysis goals of comparing social bonding rating data, rather than the acoustic recordings. Thus, it is possible that some participants from that study will pass those outcome-neutral criteria, but fail to provide reliable audio data (e.g., if the audio fails to record due to a technical glitch). It is also possible that some participants could fail their outcome-neutral controls (e.g., failing to submit the Qualtrics survey) but still provide useable audio recordings for this Programmatic protocol. Therefore, while our new protocol relies on audio recordings collected by Savage et al., these audio recordings will be subject to the following separate set of outcome-neutral inclusion criteria to ensure the recordings are of sufficient quality, duration, and reliability that they can be reliably used for our confirmatory hypothesis testing comparing acoustic features of singing vs. speaking.**

- 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.

We have added clarification as follows:

> **To be analysable, audio recordings must meet minimum standards of quality, such that our three confirmatory dependent variables (pitch height, temporal rate, and pitch stability) can be reliably measured (i.e., at least 10 units of matched singing/speaking whose fundamental frequency can be extracted; see Inclusion/Exclusion Criteria and Fig. 2 simulation below). This means they need to be recorded accurately with low enough noise and high enough quality that fundamental pitch can be automatically extracted using the pYIN algorithm[41], and the units (syllables or notes) can be clearly determined.**

- 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/

We have added an analysis of inter-rater reliability as follows:

> **Inter-rater reliability: Before annotating audio, each coder will watch the training tutorial video (Appendix S1). We will measure inter-rater reliability (IRR) following Ozaki et al. (2024) by having author Jia independently re-annotate onsets of singing and speaking from one randomly selected participant from each Stage 2 report. Like Ozaki et al., Jia will be blind to the specific onset timings annotated by the original coder, but will have access to their segmented texts (since otherwise Jia will not know the correct way to segment acoustic units such as syllables/notes spoken/sung in languages she does not speak). For reference, Ozaki et al. found "strong intraclass correlations (>0.99)" when using this method to compare 10s excerpts of singing vs speaking from 8 individuals randomly selected from the full sample of 75 individuals. Any sites with intraclass correlations of less than 0.6 (a typical threshold for distinguishing between "fair" and "good" reliability) will be independently re-checked by Savage for another randomly selected song. If this is also less than 0.6, then all songs from that site will be checked and re-annotated until they achieve coefficients of at least 0.6.**

- 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.

We have expanded the "Statistical analysis" section by adding the following bolded text:

Statistical analysis:

We will follow essentially the same analysis methods as Ozaki et al. using a meta-analysis framework to compare effect sizes from each within-participant singing vs. speaking comparison across many different participants. The main differences are:

1) we are only testing three hypotheses (pitch height, temporal rate, and pitch stability) rather than Ozaki et al.'s six
2) Each site will test whether the hypotheses replicates for its own language/society, rather than comparing across many different languages simultaneously as Ozaki et al. did (though we will also run the cross-linguistic comparison for the final meta-analysis of all 26 languages/cultures)

The full analysis plan is adapted from Ozaki et al. as follows:

We use null hypothesis testing to test whether the effect size of the difference between song and speech for a given feature is null. There are various ways to quantify the statistical difference or similarity (e.g., Kullbak-Leibler divergence, Jensen-Shannon divergence, Earth mover's distance, energy distance, Ln norm, Kolmogorov-Smirnov statistic). Here we focus on effect sizes to facilitate interpretation of the magnitudes of differences.

Since our main interest lies in the identification of whether three features - pitch height, pitch stability, and temporal rate - demonstrate differences between song and speech, we perform the within-participant comparison of these features between the pairs of singing and speech, using the alternating singing and conversation conditions as proxies for singing and speech, respectively (comparisons with synchronised singing and synchronised recitation are reserved for exploratory analyses). Terms in the computed difference scores are arranged so that for our predicted differences (H1-H3), a positive value indicates a difference in the predicted direction [cf. Fig. 5].

Evaluation of difference in the magnitude of each feature is performed with nonparametric relative effects[45] which is also known as stochastic superiority[46] or probability-based measure of effect size[47]. This measure is a nonparametric two-sample statistics and allows us to investigate the statistical properties of a wide variety of data in a unified way.

We apply the meta-analysis framework to synthesize the effect size across recordings to make statistical inference for each hypothesis [see Fig. 8 in Ozaki et al. for graphic overview]. In this case, the study sample size corresponds to the number of data points of the feature in a recording and the number of studies corresponds to the number of participants. We use Gaussian random-effects models[48,49], and we frame our hypotheses as the inference of the mean parameter of Gaussian random-effects models, which indicates the population effect size, as follows:

The Gaussian random-effects model used in meta-analysis is[48,49]:

$$Y_i|\theta_i \sim \mathcal{N}(\theta_i, \sigma_i^2),\ \theta_i \sim \mathcal{N}(\mu_0, \tau^2),\ i = 1, \ldots, K$$

Y_i is the effect size (or summary statistics) from ith study, θ_i is the study-specific population effect size, σ_i^2 is the variance of ith effect size estimate (e.g. standard error of estimate) which is also called the within-study variance, μ_0 is the population effect size, τ^2 is the between-study variance, and K is the number of studies. In our study, Y_i is the relative effect and σ_i^2 is its variance estimator[45]. In addition, the term "studies" usually used in meta-analysis corresponds to recording sets. This model can also be written as

$$Y_i \sim N(\mu_0, \sigma_i^2 + \tau^2),\ i = 1, \ldots, K$$

Our null hypotheses for the features predicted showing difference is that the true effect size is zero (i.e. relative effects of 0.5). We test three features, and thus test three null hypotheses.

Since we test multiple hypotheses, we will use the false discovery rate method with the Benjamini-Hochberg step-up procedure[50] to decide on the rejection of the null hypotheses. We define the alpha level as 0.05. We test whether the endpoints of the confidence interval of the mean parameter of the Gaussian random-effects model are larger than 0.5. We use the exact confidence interval proposed by Liu et al.[49] and Wang and Tian[51] to construct the confidence interval.

For the equivalence testing, we first estimate the mean parameter (i.e., overall treatment effect) with the exact confidence interval (98, 100) and the between-study variance with the DerSimonian-Laird estimator[52]. Since Gaussian random- effects models can be considered Gaussian mixture models having the same mean parameter, the overall variance parameter can be obtained by averaging the sum of the estimated between-study variance and the within-study variance. Then, we plug the mean parameter and overall variance into Romano's[53] shrinking alternative parameter space method to test whether the population mean is within the equivalence region as stated in Table 1 (i.e., relative effects of 0.39 and 0.61).

- 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).

Thanks for catching this omission so quickly after we submitted. We extensively updated the GitHub readme file (https://github.com/comp-music-lab/manyvoices3/blob/main/README.md) on Feb 26 before any reviewers submitted their reviews.

Review by Nai Ding, 27 Feb 2025 05:04

The study is a well motivated extension of Ozaki et al. (2024). It extends the Ozaki et al. study by adding more samples. The writing is clear and the analyses are straightforward.

We appreciate your enthusiasm (and your early work reviewing Ozaki et al. 2024 for PCI-RR).

I had the same concern I raised for Ozaki et al. (2024) - That is "inter-onset interval" is a highly ambiguous word. When taking about inter-onset interval for speech, one may think about inter-syllable-onset interval, inter-word-onset interval, inter-phoneme-onset interval, etc. The same applies for music. The term is not even defined in the draft, but even if it's defined the readers should be reminded, e.g., in the abstract and conclusion, about what kind of intervals are being considered here.

We agree this needs clarification. We have added this sentence to the abstract:

> **For each site, we will replicate Ozaki et al.'s analyses for their three key features hypothesised to differ between song and speech: 1) pitch height (f_0); 2) temporal rate (inter-onset interval of acoustic units [e.g., syllables/moras/notes]); 3) pitch stability ($-|\Delta f_0|$).**

We have also added clarification in the Registered Report design planner Table 1 ("onsets are based on acoustic units corresponding to syllables or notes in English; see Fig. 4") and added a figure (Figure 4) and clarifying text addressing this and related points below by Dr Moscoso del Prado Martin below (copied below after Dr Moscoso del Prado Martin's comment).

Review by Fermin Moscoso del Prado Martin, 04 Mar 2025 12:21

I find this is overall a well-designed study with a clear rationale. Methodologically, overall the study is clear and the statistical analyses planned are --in general-- adequate (some caveats below).

We appreciate your enthusiasm.

The methodological aspects that I think require further attention are the following:

1.- Choice of "acoustic units". The authors propose limiting the analyses to a fixed number of acoustic units. If I understood correctly these are syllables. This might be problematic. Whereas the syllable is indeed a crucial unit in many languages (referred to as syllable-timed languages in linguistics, examples in this study's set would be Spanish or Mandarin, among others), it is not so for all languages. Other languages only use the spacing between *stressed* syllables as main

units (referred to as "stress-timed languages" in linguistics, and in their set these would include English or Farsi), and finally, other languages do not use syllables at all as their timing units, relying instead on "moras" (referred to as "mora-based" languages, such as Japanese). In view of this typological difference I would suggest than rather than syllables (which are not acoustic units in some of their languages) I would fix either a time duration, or use some other units such as phones or words. This is particularly relevant for any measures of speed.

We agree that we cannot universally apply the concept of syllable as acoustic unit across languages. We have added the following figure and explanation:

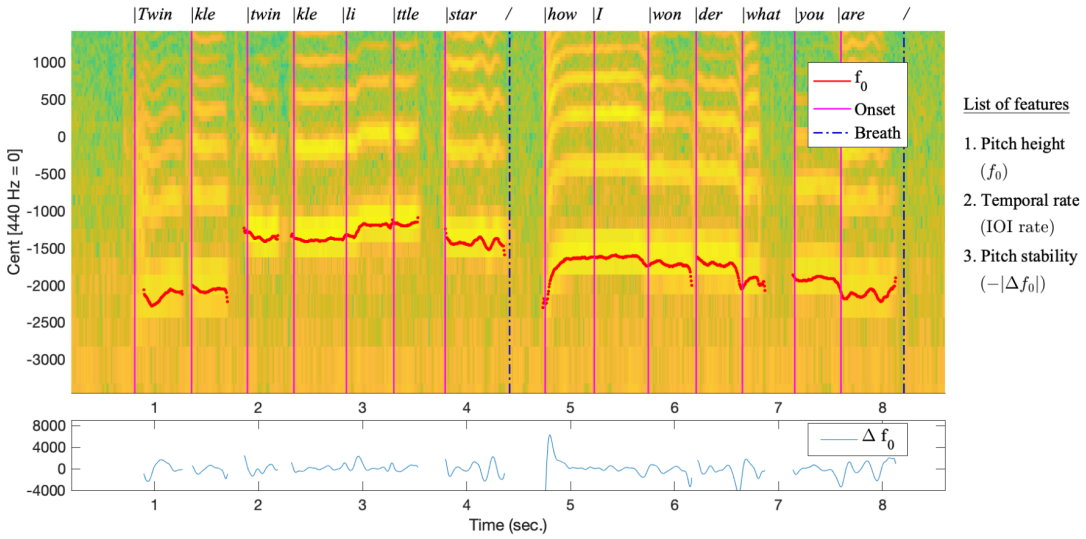

Figure 4. Schematic illustration of the three features analysed for confirmatory analyses, using a recording of author Savage singing the first two phrases of "Twinkle Twinkle Little Star" as an example. This figure is identical to Figure 3 in Ozaki et al., but only shows the three features proposed to test here of pitch height, temporal rate, and pitch stability. Onset annotations used to calculate Inter-Onset Intervals (IOIs) are based on the segmented texts displayed at the top of the spectrogram (breaths are excluded from IOI calculations). For this English song, these onsets correspond to syllables, which also correspond to sung notes, as Twinkle Twinkle uses one note per syllable. However, the choice of acoustic units can vary depending on the language and song[42]. For example, in Japanese it typically corresponds to a "mora" (e.g., みんな = mi|n|na) and songs often use multiple notes per syllable[43]. Following Ozaki et al.[2], the appropriate segmentation unit for each language is chosen by the lead researcher analysing that language (who is also a speaker of the language). In most cases, these units will be syllables or moras for speaking and notes for singing. Note that our analyses are intended to address speaking/singing rate, so higher- or lower-level units such as stressed syllables, metric downbeats, or phonemes[8,44] are not the focus of our confirmatory analyses.

2.- Choice of songs. According to the design, it seems like the choice of songs that the subjects will produce will be made by the experimenters. This could be subject to a confirmation bias. The chosen songs could --inadvertedly-- be chosen to have higher pitches. May be allowing

subjects to choose their own songs, or choosing the songs according to some explicit criterion could help on this.

We have clarified the song choice criteria as follows. We also address the question of confirmation bias below in response to a similar point by Dr Sadakata:

> **Each site in Savage et al.1 will recruit 15-30 participants and choose its own song (cf. Table S1 from ref. 1) and conversation prompt using the following criteria. Note that the need to recruit participants to sing together in groups means it is not feasible to allow each participant to choose their own song as Ozaki et al. did:**
> **…**
>
> ***Song selection criteria***
>
> ***Each site has chosen a song that would be appropriate for their language/culture. The criteria for choosing a song were:***
>
> ***-lyrics are mostly in the same language that participants will use for their group conversation (some lyrics in other languages or meaningless vocables like "la la" are acceptable, but should not make up the majority of the song)***
>
> ***-should be easy for most potential participants from that society to sing together in synchrony (e.g., unison, homophony) with karaoke-style pre-recorded instrumental accompaniment without needing to practise ahead of time. If possible, this should be in the form of a karaoke-style video with plain background and lyrics that appear in real-time to help the participants to sing at the right time, with no guide melody (e.g., https://youtu.be/OhRUYf_yn_s?si=eL4mt_-utRwqrFMj&t=10). If pre-recorded instrumental accompaniment would not be appropriate for a given site/society, an a cappella (unaccompanied) song may be chosen instead.***
>
> ***-should be the kind of song that would be appropriate to sing by young adults who don't already know each other as a short "ice-breaker" exercise. As such, songs that might easily become awkward, embarrassing, or offensive should be avoided (e.g., children's songs, songs with polarising content or associations such as national anthems or religious songs). However, these factors may vary from site to site (e.g., for some communities a national anthem or religious song might be the best choice, while in others it might be the worst). The experimenters from each site should interpret this on the basis of their own local knowledge.***
>
> ***-the song should take between 2-3 minutes to sing (you are welcome to modify the number of verses/choruses (including repeating the song) to make this happen***
>
> ***-if the song has instrumental interludes/introductions/outros, these should not be longer than 1 minute total and there should still be 2-3 minutes of singing time not including these instrumental sections.***

3.- On the analysis side of things, not much discussion is given on how they would deal with different languages providing different results. It seems like they consider a binary result, but the truth may well be that some languages show that distinction, some may show the opposite, and

some might be unclear. How will such cases be dealt with. Will they explore the socio-cultural and linguistic typological differences that may lead to those?

We have clarified this point in the "Hypotheses" section as follows:

> **Since this is a Programmatic Registered Report where one Stage 1 protocol will result in multiple Stage 2 outputs, it is possible that different Stage 2 outputs will produce different results for different languages. This will allow us to evaluate criticisms that global analyses of cross-cultural trends fail to address the importance of internal diversity ("*How many exceptions are researchers willing to ignore?*"[28]).**
>
> **Except for the Stage 2 output combining all studies (#27 in Table 2), each Stage 2 will focus its confirmatory analyses on the results of its own analysis of its own focus language. #27 will replicate Ozaki et al.'s cross-linguistic meta-analysis approach to analyse average trends across all languages, which can be compared with the results of each individual Stage 2 reports #1-26 to achieve a much broader evaluation of the cross-linguistic replicability and generalisability of Ozaki et al.'s original results. Comparison of specific differences between languages will be reserved for exploratory analysis (since statistical power for such comparisons will be limited by the relatively small sample size of n=15-30 participants per language).**

Review by Makiko Sadakata, 28 Feb 2025 13:17

Dear Editor and Authors,

This is a well-designed and thoughtfully structured study that builds on previous research. The project has clear research questions, a solid methodological foundation, and well-motivated hypotheses. Its collaborative nature adds further value. I have a few questions, which are intended to enhance clarity and strengthen the study's design rather than to criticize it.

Structure
The proposed structure is well-coordinated. I understand that (1) since each site tests the hypotheses independently and reports its own findings, this avoids issues of overcrowded reporting, and (2) the meta-study synthesizes broader trends, adding value without redundancy.

We appreciate your enthusiasm.

However, because this structure is new to me, I have a few clarification questions.

While the Stage 1 report ensures 'in principle acceptance' for the three pre-registered hypotheses at each site, it does not extend to any additional hypotheses that sites may introduce. Since these additional hypotheses do not alter the core findings related to the pre-registered questions, I do not see this as a major concern. However, a brief conceptual clarification on how the scientific rigor of these additional hypotheses is maintained—whether through a standardized process or left to individual sites—would be appreciated. This would also help clarify how these hypotheses will be evaluated during the review process. If their assessment is entirely left to the reviewers of each separate paper, it would be useful to make this explicit so that expectations are clear.

We have clarified as follows:

To ensure maximal consistency across Stage 2 reports, all Stage 2 reports will restrict their confirmatory analyses and statistical hypothesis testing to only these three hypotheses. They are welcome and encouraged to explore additional analyses, but must ensure these conform to PCI-RR's Stage 2 criterion 2D[39]:

> ***2D. Where applicable, whether any unregistered exploratory analyses are justified, methodologically sound, and informative***

The paper states that there will be "up to" 27 individual reports, which suggests that some sites may not publish their findings independently. If some sites do not publish, will their data still be included in the meta-analysis? If so, how will their data be handled to maintain quality control?

Given the scale of coordination, it would also be helpful to consider potential unexpected scenarios and corresponding action plans. For example, in Savage et al. (2025), if a certain percentage of sites failed to deliver data, a contingency plan was in place to supplement missing data. Would a similar stepwise plan be considered here to ensure the meta-analysis remains robust even if some sites do not complete their reports or data collection on time?

Good points. We have added the following criteria and explanation:

> **Note that the original studies this Programmatic Registered Reports replicates and extends had very different minimum sample size requirements: Ozaki et al. (2024) specified a minimum sample size of 60 participants, while Savage et al. (2025) required a minimum sample size of 450 participants total (minimum of 30 sites, each with a minimum of 15 participants). For this Programmatic report, Stage 2 reports #1-26 will rely on acoustic data from the subset of sites from Savage et al., and thus also have a minimum of 15 participants each. However, it is possible that in some sites the number of participants with analyseable singing and speaking audio recordings may be fewer than the number of participants (e.g., if a participant does not have a chance to speak during the conversation condition).**
>
> **We will specify a minimum of 10 participants per site for the 26 proposed single-site Stage 2 reports. For the meta-analysis (#27), it is likely that some of the 26 proposed sites will not be able to complete their Stage 2 Reports within. However, a meta-analysis of even a small number of sites would still be valuable, meeting criteria such as *Advances in Methods and Practices in Psychological Science*'s "Registered Replication Reports (RRRs)", which require "direct (i.e. close) replications in any area of psychology that involve coordination between at least three (but preferably more) independent teams of researchers" (https://rr.peercommunityin.org/about/pci_rr_friendly_journals#h_91557356867416524390 66888). For consistency with this and with Ozaki et al.'s original minimum sample size of 60 participants for cross-linguistic meta-analysis, we will plan to continue the meta-analysis Stage 2 Report (#27 in Table 2) even with as few as 60 participants worth of data from as few as 3 sites are collected, analysed, and published as Stage 2 reports (i.e., minimum sample size for the meta-analysis of 60 participants from a minimum of 3 languages).**
>
> **…**
>
> **-Minimum sample size for Stage 2 reports #1-26: 10 participants**

-Minimum sample size for Stage 2 report #27: 60 participants speaking at least 3 different languages
-For the meta-analysis (#27) all useable data from Stage 2 reports #1-26 collected and analysed within 18 months after In Principle Acceptance will be included
…
-If any sites choose to withdraw (i.e., not to publish a Stage 2 report), their data will also not be included in the meta-analysis confirmatory analyses in #27. In such cases, the meta-analysis will report the reasons for withdrawal (e.g., lack of time to analyse data or write up analyses; researcher graduating/changing jobs; data not meeting inclusion criteria standards) and describe how much, if any, of the data were collected/analysed before withdrawal, summarising any preliminary results if they exist. Note that we cannot commit to analysing all data if sites withdraw because our proposed acoustic analyses require time-consuming manual annotation by researchers with knowledge of the local language/music. However, we commit to not making decisions about whether or not to withdraw based on how these affect our conclusions.

The original data collection plan in Savage (2025) is set to be implemented at 57 (or 60? There seems to be a discrepancy between the two papers—please check) sites, while this study will use data from 26 of those sites.

3 additional sites joined the collaboration during the two months of review preceding In Principle Acceptance. We have updated the numbers of sites in Savage et al. (2025) from 57 to 60 (notifying the Recommender). This does not affect our specified minimum number of sites, which remains at 30).

To enhance transparency, could you clarify whether the selection of these 26 sites was determined before data collection began? If the selection took place after data collection but before transcription, could this introduce potential selection bias?

We had not begun Stage 2 data collection at any sites before submitting this Programmatic Registered Report. We have added the following clarification:

> **The selection of these 26 sites was determined before any Stage 2 data collection began for Savage et al.[1]. All 60 research sites were invited to participate. Inclusion in this Programmatic Registered Report depended only on the interest and availability of researchers at each site. In particular, they had to be willing to wait to begin data collection until this Programmatic Registered Report also receives In Principle Acceptance from PCI-RR to ensure maximum bias control (Level 6: "*No part of the data or evidence that will be used to answer the research question yet exists and no part will be generated until after IPA [In Principle Acceptance]*"[39]).**

Methods (Discussion point)
If participants first engage in a karaoke-style singing task, could this prime them to a specific tempo, influencing their subsequent monophonic singing? The study examines tempo differences between speech and song. In this context, is it important to consider whether the methodology captures the natural singing tempo? The karaoke accompaniment might shape the singing tempo rather than reflect a spontaneous pace. While I understand that data collection may already be in

progress and the protocol cannot be changed as it is pre-registered, it would be valuable to discuss this as a potential impact.

Excellent point. Since we had not yet begun Stage 2 data collection for Savage et al. (2025), we modified that protocol to counter-balance condition order, as well as adding/replacing the following bold/strikethrough texts to the current protocol:

> **Randomisation: At each site, the 15-30 participants will be randomly assigned into one of three groups. Each group completes the same four conditions (conversation, monophonic singing, unison singing, lyric recitation) but in different orders. When the (unaccompanied) monophonic singing condition follows the unison singing accompanied by karaoke-style accompaniment, participants may be influenced by having just heard and sung at the key and tempo matching this accompaniment. Likewise, it is possible that people may sing/speak differently depending on whether they have a conversation before or after singing. For these reasons, Savage et al. counter-balanced the order of conditions in the three participant groups as follows, enabling exploratory analyses of potential order effects:**
>
> ***-Group 1: 1) conversation, 2) monophonic (alternating) singing, 3) unison singing, 4) lyric recitation***
>
> ***-Group 2: 1) unison singing, 2) lyric recitation, 3) monophonic (alternating) singing, 4) conversation,***
>
> ***-Group 3: 1) lyric recitation, 2) unison singing, 3) conversation, 4) monophonic (alternating) singing***
>
> ~~In two of these groups, the conversation will be recorded before the alternating singing, while in one group the alternating singing will be recorded before the conversation. (This order is unlikely to affect results, but this can be investigated in exploratory analyses.)~~
>
> …
>
> These possibilities (and others, **such as potential order effects described in the "Randomisation" section above**) may be worth addressing in the Discussion section and Exploratory Analysis sections of the resulting Stage 2 reports.

Sincerely, Makiko Sadakata

Savage, P. E. et al. Does synchronised singing enhance social bonding more than speaking does? A global experimental Stage 1 Registered Report [In Principle Accepted]. Peer Community Regist. Rep. (2025) doi:10.31234/osf.io/pv3m9.