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

29th November, 2024

RE: ***Does synchronised singing enhance social bonding more than speaking does? A global experimental Stage 1 Registered Report*** **(https://doi.org/10.31234/osf.io/pv3m9)**

Dear Dr. Moore,

We appreciate your invitation to revise and resubmit our Stage 1 Registered Report protocol based on the constructive comments of the four Reviewers. We are grateful for the chance to use their feedback to modify our planned experimental protocol to enable us to reach stronger conclusions following our eventual Stage 2 data collection and analysis. In particular, we have made the following key changes:

1) changed topic of the conversation condition from the song lyrics to a generic ice-breaker question not linked to song lyrics or music

2) modified and added song selection criteria, modifying proposed songs for some sites accordingly, with all conditions now taking between 2-3 minutes and including karaoke-style accompaniment

3) differentiated specific hypothesized predictions for singing vs. sequential speaking (H2a) and singing vs. synchronous recitation (H2b) and broke down the main research question into two separate questions/hypothesis sets: 1) does singing enhance social bonding? 2) does singing enhance social bonding more than speaking does?

4) added two self-report questions to the previous four used to create the "social bonding" dependent variable (for confirmatory analyses) and a public goods game measure of behavioral prosociality at all sites (for exploratory analysis purposes)

5) added video monitoring of experiments to ensure instruction compliance

In addition to these major changes to the protocol, we have also changed the manuscript to address other points, particularly regarding clarifying terminology in the title, abstract, Registered Report Design Table, and throughout the manuscript (e.g., changing "cooperation" to "social bonding" and "language" to "speech"). Among other changes, we have also changed our Bayes Factor threshold from 10 to 3 (approximately equivalent to p<.05; ref. 2) and updated our citations appropriately now that we no longer aim to publish in *Nature* where BF>10 is required. 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 (Katherine Moore):

Thank you for submitting your pre-registration report titled "Does synchronised singing enhance cooperation more than speaking does? A global experimental Stage 1 Registered Report" Four expert reviewers have read your proposal and provided feedback. Collectively, the reviewers praised the study. They believe the work is on an important topic and is likely to have a very positive impact on the literature. They are especially impressed with the scale of the collaboration and cross-cultural investigation. Of course, for that reason, all reviewers also believe it's important to have the best possible design to maximize the fruits of this labor. They have each offered helpful suggestions on how to improve the study.

The comments include suggestions and concerns about the literature review and theoretical context as well as with the design of the study. Some of the reviewers have provided similar concerns. For example, Drs. Hannon and Marin both suggested videotaping the session to evaluate participants' compliance with the task instructions, or perhaps going a step further to ensure compliance in the moment.

We have implemented the helpful suggestion of ensuring compliance via video monitoring.

Several reviewers brought up concerns about equivalency between the speech and song tasks and across sites. They suggested some possibilities to address this, such as requiring participants to repeat short songs up to a particular duration, or to choose poems instead of songs for the speech task.

We have addressed this by unifying the duration of the singing and speaking conditions to be 2-3 minutes across all sites (allowing for repetition of short songs) and by changing the conversation task from being about song lyrics to a generic ice-breaker question not related to song lyrics or music.

One important issue for the authors to consider is weighing external (or ecological) vs. internal validity. Drs. Hannon and Goritz both suggest using another measurement of cooperation other

than self-report, and provide some suggestions. Drs. Brandon and Marin both make a suggestion for improving terminology.

We have added the suggested public goods game as an exploratory measure at all sites, and changed key terminology throughout (e.g., using "social bonding" instead of "cooperation", "speech" instead of "language").

These are some examples, but most reviewer suggestions are unique. Please address all of the reviewer comments in a revision of your report. It is not necessary to take every piece of advice from the reviewers, but you should address their concerns (and why you disagreed, if applicable) in a response letter.

Thank you for your submission and I look forward to seeing your revision!

*by **Katherine Moore**, 27 Oct 2024 03:41*
Manuscript: ***https://doi.org/10.31234/osf.io/pv3m9***
version: 6

Review by Melissa Brandon, 16 Oct 2024 04:18

Dear Editor and Authors,

Overall, this is a very exciting study proposal that has a well thought out plan to collect data across multiple sites around the world. The research question is clear and will provide useful knowledge to the field about music's impact on social cooperation with additional controls missing from past studies. There is a logical plan for the primary data analysis. I do have a few questions and suggestions for clarity listed below by line numbers. I look forward to seeing the study results in the future.

Sincerely,

Melissa Brandon, Ph.D.

Thank you for this positive and constructive review!

Questions or comments:

Lines 78 & 79: Sentence is worded awkwardly. Can you use the word conversation over language or does that deter from the point of comparing language and music?

Good point: in response to this point and a related one by Dr Göritz, we have changed and condensed this phrase to read:

> *Nevertheless, some remain skeptical that music specifically causes* ***social bonding***

Line 191: Can you say more about the measure of cooperation? Is this a normed survey and if so in how many cultures or languages? Or how is the survey being adapted across languages and cultures? Interested to know how comparable this measure is across different cultures and languages. Add this either here or around line 243 or 329. But need more information on the measure and its adaptability across sites.

We have added two more variables, along with more explanation about the rationale for choosing them to (what was previously) line 329 in response to related critiques by Dr Marin (see below). We have also added the following clarification:

> ***"The final (In Principle Accepted) version of the Qualtrics survey and instruction video texts will be translated/adapted to the language and song lyrics of each site by the authors responsible for data collection at that site (see Table S1)."***

We have also added an example of what the translated version looks like to Figure 4 (note that we will wait to fully update the Qualtrics survey and its translation with the newly added variables until the final protocol is nearly fixed):

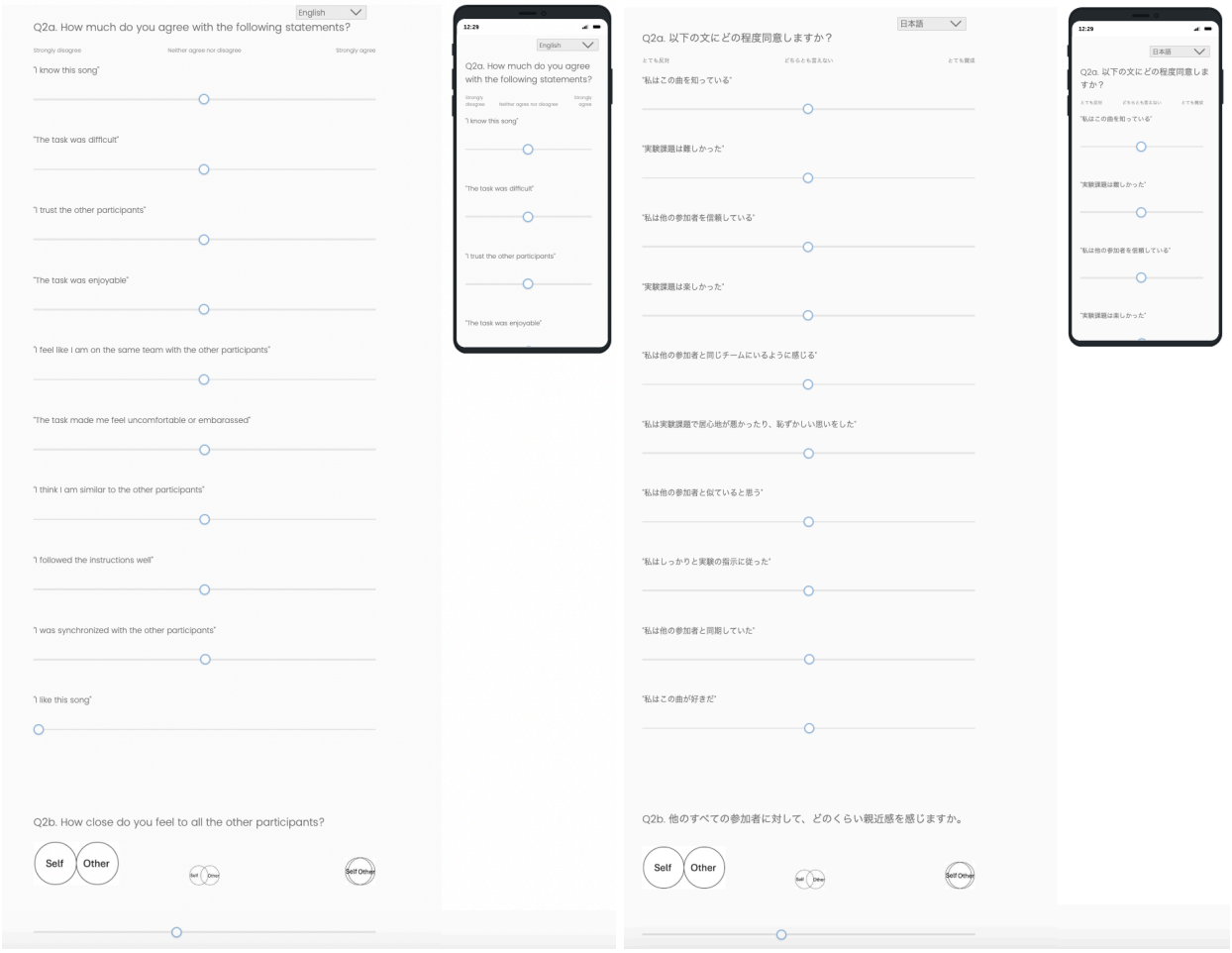

Figure 4. Screenshot of the setup of key variables using 0-100 sliding scales via Qualtrics, showing the English version (left) and an example translated version in Japanese (right)....

Line 222: The post-interaction assessment of corporation needs to be stated here in the method plan and any info on your other post-test questionnaires. Do they all happen after the interactions? Given that the cooperation measure post interaction is your main source of data it should be stated in the procedure plan.

We have added the following text:

> **2) Post-interaction: The same variables from the pre-interaction phase (0) will be collected again.**

Projector text: "Please fill in the next page of the survey on your device now."

3) Demographic variables and debriefings: Additional demographic variables will be collected for exploratory analysis, along with a brief debriefing text.

Projector text: "Follow-Up Questionnaire: Please fill out the remaining page of the survey on your device. Thank you! Feel free to leave whenever you finish, even if the other participants are not done."

Line 258: Further clarify "able to sing song with lyrics." Will this self-select to only people willing to publicly sing? That could skew the sample to people with music experience. On the other side are you trying to ensure no one with amusia is in the study? I am asking are you looking for a sample with all levels of music experience or trying to exclude anyone?

We have added the following text after the inclusion criteria:

Note that, while we welcome and do not intentionally exclude individuals with limited musical/linguistic abilities or experience, our requirements for participants to be willing and able to sing a specific song and have their voices shared publicly is likely to inevitably result in selection bias in recruitment, as better singers, more extroverted people, etc. may be more likely to volunteer to participate. Our participants will therefore not represent a random subset of the broader target population(s). We will interpret our results in light of this and other limitations.

Line 277: Good clarifications of standards of when and how data is kept in case of small groups and or technical errors. Glad this was informed from pilot data.

Thanks!

Line 553: For the "Additional data types", do you have the number of sites collecting the extra measures or the goal # of participants in those additional measures? Good to add if you already know the information.

Actually, upon considering the critiques of Dr Marin below about the possibility of additional data type collection logistics affecting the main experiment, we have decided to cancel our plans to collect subsets of additional data types and focus only on collecting self-report and public goods games at all sites, modifying the manuscript accordingly.

Lines 572-578: Are all sites doing the post-experiment conditions or only some? Will the same analysis approach be used for the 3rd cooperative measure as described above or a different analysis since this is in the exploratory section, please clarify?

We have added the following clarification. Note that authors are neither expected not encouraged to pre-specify analysis plans for exploratory analyses in Stage 1 protocols so we have not specified analyses further here:

> **Data from these post-experiment exploratory conditions will be collected from all sites. However, if for any reason we fail to collect usable data from these post-experiment exploratory conditions but do collect usable data for the primary confirmatory analyses, we will still include these data in confirmatory analyses.**

In looking at Table S1 there are differences in the values of compensation participants will be paid. I noticed because values of the different locations in the UK are very different amounts. I know this is determined by site, but it may be another variable to examine. If the values are drastically different it could affect motivation to participate and potentially the feelings of cooperations. If there is a similar difference in compensation across sites in your pilot data, it may be worth examining if it is an explanatory variable. This could help you plan for how to control for this difference in your study (line364).

Thanks for this suggestion, we have added this along with several other potential site-level exploratory variables as follows:

> **Site-level variables: The following additional exploratory variables will be investigated to explore potential factors affecting all participants at a site:**

1) **Singing/speaking language**
2) **Participant compensation (e.g., raw amount [in USD equivalent], relative Purchasing Power Parity, etc.)**
3) **Musical/acoustic features of chosen songs (e.g., tempo [bpm], pitch height [Hz], emotional valence [0-100 negative-positive subjective rating by researcher team who chose the song], etc.)**

Review by Erin Hannon, 15 Oct 2024 15:33

The proposed study asks a valid and novel question about the effects of synchronous musical activity (singing) on cooperation, comparing this with sequential and synchronous (chanting) speech conditions relative to a pre-intervention baseline. The study will have a large, multi-site, multicultural sample and thus address questions about music and cooperation in a much more diverse sample than has previously been tested. There are excellent controls and procedures in place (experimenters are blind to condition, stimulus decisions are left to local experts), and crucial additional variables are being tracked (the extent to which participants knew each other beforehand, music training, etc).

Thanks for this positive and constructive review!

At this stage, I have two minor concerns/questions and one more substantive concern about the proposed study.

1. As I understand it, the experiment will be run in a room without an experimenter and instructions will be given over video. This is to ensure experimenter blindness, and it seems fine, except as far as I can tell there is no procedure in place to measure the extent to which participants followed instructions and did the task as instructed. This is especially important because they are not supposed to interact prior to the singing/speaking, but it also seems important if some groups do a better job than others singing in synchrony etc.. Most of the between-group variation appears to be accounted for through self-report. It seems like it would be a relatively simple step to video-record sessions and later check that participants followed instructions, at least for those sites that can do so. If such procedures are already in place, they should be described in the document, along with details about how compliance would be evaluated.

Thanks for this suggestion - we have implemented it as follows (in the "Exploratory analyses" section):

> **Cohort-level variables: Each experiment will be monitored in real time by the local experimenters via Zoom video, where the instruction video will also be shown using screen share (the experimenters' video and audio will be muted). These videos will not be published, but will be used by the experimenters to monitor compliance and allow them to intervene if subjects misunderstand instructions, in case of an emergency, etc. After each experiment, experimenters will rate the following variables. Note that these ratings can only be done at the cohort-level and cannot be linked directly to**

individual participants because individual participant surveys are done anonymously via Qualtrics.

1) **Experiment date**
2) **Experiment start time**
3) **Experiment location**
4) **Experimenter name**
5) **Number of participants (NB: This may vary from the number of Qualtrics responses - for example, if one participant from a group of 10 fails to complete the Qualtrics survey, only 9 responses will appear but the "Number of participants" for that cohort is 10.)**
6) **Instruction compliance (0-100). NB: This will be rated after the experiment but before observing the Qualtrics survey. All participants from groups where "Instruction compliance" is judged unacceptable by the experimenter (<25 out of 100) will be excluded and experimenters will re-recruit a replacement group of participants.**

2. This is perhaps just a matter of clarification, but I did not understand the following: "Although the (synchonised) singing condition and the (simultaneous) conversation conditions differ in the presence of both singing and synchrony, we cannot measure the effect of these factors separately through the comparison of only these two conditions. Therefore, we model the combination of these effects as a single effect, which we name synchronous singing." Does this just meant that one variable will dichotomously code for speaking vs. singing, and another variable will code for synchornous singing vs. not synchronous singing? If so this could be made clearer.

Thank you for bringing up this point. As we have decided to add different predictions for singing vs. lyric recitation and singing vs. conversation (at Dr Marin's suggestion below), we have updated the sentences as follows:

Although the (synchronised) singing condition, the (synchronised) recitation, and the (sequential) conversation conditions differ in both modality and synchrony, evaluating these effects individually brings complexity to designing decision rules for determining which features should be significant in rejecting the null hypotheses H0a and H0b. Therefore, we choose to model the combination of these effects as a single effect, resulting in the use of indicator variables for the three conditions.

3. My biggest concern is regarding the construct validity of the dependent measure(s). The research proposes to measure cooperation with only 4 self-report responses. I realize that other studies examining effects of synchronous activity on cooperation have used similar measures, however given the magnitude and potential impact of this project, it seems like a missed opportunity to not also have more direct measures of cooperation. If participants figure out what the researchers are measuring (which is likely) this could give rise to demand effects-- yes, this would affect all three conditions, however the pattern of results could reflect more about what they believe about the effects of various activities on cooperation rather than how they actually feel towards other participants after doing the activities. Why not also ask them to make a decision about sharing a resource with the group? There are many simple "games" in the behavioral economics literature that could be easily implemented in Qualtrics. The down side would be ensuring all sites are able to offer some sort of resource of incentive. If the authors disagree, then perhaps they could add some discussion acknowledging the limitations of only using a self-report measure of the outcome variable.

Thanks for this suggestion, shared with other reviewers. We have added two more self-report variables, and also added the behavioral economics game measure at all sites for exploratory analyses, along with the explanatory text below:

> **Public goods game:**
>
> **Previous studies exploring links between synchrony and social bonding/prosociality have used a variety of proxies, including self-reported attitudinal prosociality and behavioral measures via behavioral economic games (e.g., stag hunt, public goods game)6,26,27. We have chosen to focus our confirmatory analyses on self-reported variables rather than behavioral economics games, for the following reasons:**
>
> 1) **Meta-analyses suggest equivocal results with behavioral economics game measures.6,26**
> 2) **Compared to behavioral measures, self-report measures are often more reliable, practical, flexible, and inclusive (especially important for our multi-site cross-cultural design), without being necessarily worse regarding expectancy effects59.**
> 3) **Our pilot experiments suggested concerns with possible ceiling effects (a majority of participants in pilot experiments chose to contribute the maximum possible amount).**
> 4) **Behavioral economics game measures can require careful calibration of the monetary incentives via iterated pilot experiments to capture the intended effects. This is challenging even for one or a few sites, and unfeasible for our set of 57 sites spanning diverse languages and economies.**

We will collect a behavioral economic measure, but will limit it to exploratory analyses (see "Public goods game" section for details), since any discrepancies between self-report and behavioral results will be difficult to interpret conclusively.

For all sites, we will also collect a measure of a monetary contribution in a cooperative public goods game in addition to subjective self-report ratings, using the following scenario: "Imagine that all participants can anonymously contribute some of their payment to a pool of money that will be multiplied by 1.5 and divided equally among the participants. The more you contribute, the more all participants will receive on average, but the less you contribute the more you personally will receive in the end. So, if everyone in a group of 5 contributes 50% of their payment (i.e., $5), each person in the group would get $12.50 instead of $10 [amount/currency to be adapted based on each site's payment schedule]. If, however, the others all contribute 50% but you contribute 0%, the others would receive $11 total but you would receive $16. No one will know how much any other individual chose to contribute, only the total amount.

How much would you realistically want to contribute to the shared pool (from 0%-100%)?"

Note that we have chosen to open this question with "Imagine that…" in order to maximise comparability across sites given that not all countries/institutions/labs allow paying real money to participants in varying amounts or deceiving participants into falsely believing they will be paid real money.

Review by Manuela Maria Marin, 01 Oct 2024 19:51

I commend the authors' cross-cultural initiative to study the effects of singing and speech on social bonding. Their efforts may lead to a valuable contribution to the existing literature. I hope these comments will be helpful as the authors continue to work on their project.

Thanks for the positive, constructive, and detailed review!

Introduction:

a) l. 59-66: This section misleadingly reads as if Darwin had himself no theory about the origins of music, but in fact, he actually proposed a sexual selection hypothesis for music, which has also gained empirical support. Please add this information after l. 62, two recent reviews of the literature can be found below:

Bamford, J. S., Vigl, J., Hämäläinen, M., & Saarikallio, S. H. (2024). Love songs and serenades: a theoretical review of music and romantic relationships. Frontiers in Psychology, 15, 1302548.

Marin, M. M., & Gingras, B. (2024). How music-induced emotions affect sexual attraction: evolutionary implications. Frontiers in Psychology, 15, 1269820.

We had actually included some additional text/references (including Marin & Rathgeber 2022) on this in the previous version of the manuscript (v5: https://osf.io/download/66734e162026e9019a23e268/?version=5) but deleted it before submission due to concerns it would be too confusing since we cannot test the sexual selection hypothesis in this experiment. But upon reflection, we agree it deserves reinstating with a brief mention/citation, so have added the following sentence to the opening paragraph (including the two new references suggested):

> **Darwin speculated that musicality may have evolved via sexual selection, though this hypothesis remains controversial and difficult to test (for reviews, see 3,4,12,16–20).**

b) l. 56-108. Use of terminology: Please define and introduce the terms social bonding, social cohesion and prosocial attitudes and behavior. I would also recommend using and focusing on the concept that fits best the dependent variable of interest (prosocial attitudes or social cohesion?). I suggest using the term speech (or speaking) rather than language because speech refers to the auditory component of language and is thus more appropriate in the current research context. I know that the authors are aware of it, but testing the effects of singing (with words) and speech on prosocial attitudes is a very specific comparison that cannot easily be generalized to other forms of human music, such as joint music making involving instruments, dancing, listening to music and so on.

Great points. We have replaced "cooperation" and "attitudinal prosociality" with "social bonding" and "language" with "speech" throughout most of the paper, and added the following quoted definition when introducing our primary "social bonding" dependent variable:

> **The social bonding hypothesis defines social bonding broadly as:**
>
> **"*the formation, strengthening, and maintenance of affiliative connections ("bonds") with certain conspecifics (i.e., the set of social processes that engender the bonded relationships that underpin prosocial behavior)... we use "social bonding" as an umbrella term to encompass both bonding processes (over short and longer time***

> ***scales) and their effects. Consequently, we take "social bonding" to encompass a variety of social phenomena including social preferences, coalition formation, identity fusion, situational prosociality, and other phenomena that bring individuals together."*** **5**

c) l. 63-81: It would be useful for readers and reviewers to get an estimate of the evolutionary time-scale the authors have in mind for their social bonding hypothesis and the evolution of music and speech in humans more generally. Do the authors think that music and speech may have co-evolved among humans with a common precursor? I think it is important to get further insights into these issues because one may ask why the authors chose speech as a control condition and not some other group activities in which interpersonal synchrony also plays a role, such as group sports, working/walking together, or activities in creative arts unrelated to language. To sum up, the choice of the speech conditions should be better theoretically justified.

d) l. 72-81: Our closest evolutionary relatives are group-living animals, and unlike humans, they do not exhibit complex speech and do not show any signs of music. This is actually a valid argument against the idea that social bonding and cooperation are the (only) driving forces behind the evolution of music. Effective survival in groups neither requires a complex language nor music. Please comment.

Thanks for both of these points. As with sexual selection above, these are complicated issues, many not testable in the current design, and so we do not wish to distract by adding too much speculation (indeed, not all of our 70+ authors endorse the social bonding hypothesis). However, we agree more context would be helpful, so have extended the third paragraph and added an extended quote from the social bonding hypothesis authors to make this clearer as follows. Note that the social bonding hypothesis does not propose that social bonding and cooperation are the only driving forces behind the evolution of music:

> **Like the sexual selection hypothesis, the social bonding hypothesis is also controversial and difficult to test (for review, see refs.3,4,12,16–19 and the 60 commentaries accompanying refs.3,4). However, it does make specific predictions that can be tested in contemporary human populations, such as:**

> **"*music (including dance) is better-suited to social bonding of large, complex groups than ABMs [Ancestral Bonding Mechanisms] (grooming and laughter), language, or other non acoustic bonding mechanisms such as shared decorations or non-musical ritual behaviors (e.g., praying together without music). Music should be more effective***

***and/or efficient relative to other methods as group size and complexity increase, such that while making music in pairs might only produce a small increase in dyadic bonding relative to conversation, making music in larger, more complex groups of people (dozens or hundreds organized into differentiated sub-groups) should be <u>more effective for collective bonding than language</u>, laughter, grooming, and so on*.”4**
[emphasis added]

e) l. 133, hypotheses: I think that the authors could formulate more specific hypotheses, based on references 43 and 44, and due to the fact that they are planning to have two speech conditions (recitation vs. conversation) and a baseline condition. For example, one could write “Singing enhances prosocial attitudes more than recitation and conversation do in comparison to a baseline condition” or if the authors want to be even more specific and think that conversation should lead to the lowest levels in prosocial attitude, the hypotheses could also reflect this. If the authors make changes, please also do so in Table 1. I think that the baseline condition should be mentioned in the hypotheses. If not, please explain why.

Good suggestion, we have updated the hypotheses (and Table 1) as follows to accommodate these points and related points by Dr Göritz below:

Question 1. Does music <u>enhance</u> social bonding?
Alternative hypothesis (H1): Synchronous singing <u>enhances</u> social bonding relative to a pre-experiment baseline
Null hypothesis (H0a): Synchronous singing <u>does not enhance</u> social bonding relative to a pre-experiment baseline

Question 2. Does music enhance social bonding <u>more than speech</u> does?
Alternative hypothesis a (H2a): Relative to a pre-experiment baseline, synchronous singing enhances social bonding <u>more than conversation</u> does
Alternative hypothesis b (H2b): Relative to a pre-experiment baseline, synchronous singing enhances social bonding <u>more than synchronous recitation</u> does
Null hypothesis (H0b): Relative to a pre-experiment baseline, synchronous singing <u>does not enhance</u> social bonding <u>more than</u> sequential <u>conversation or</u> synchronous <u>recitation</u> does

f) Are there other studies than references 43 and 44 that have tested the effect of speech on social bonding? If so, they should be cited.

We have clarified that we are not aware of any other previous studies that have directly tested the social bonding effects of song vs speech as follows:

> **The one study that did directly compare music with speech reported a strong increase in self-reported trust after group singing (n=24 participants) when compared with group poetry recitation (n=24)49. The only other study we are aware of directly comparing social bonding effects of singing vs speech (published after Rennung & Göritz's meta-analysis) also reported**…

Methods

a) l. 177, group testing and group size: I think it is problematic that the planned group size varies so much (although Rennnung & Göritz (2016) did not report a group size effect), especially for the conversation condition it may make a difference whether conversation takes place between 5 or 10 people in a group. If time is limited, a large group may hinder conversation and prevent every participant from contributing, which makes it a flawed comparison. Note that the authors are aiming for group conversation to happen, and not for individual conversations between pairs. Why not make it 5 people per group across all sites? One can run more than one session per site.

We previously discussed these trade-offs with our collaborative team and concluded that 5-10 participants was the optimal size when combining theoretical and practical trade-offs. We actually would have preferred to get 10 or even much larger samples to properly test, but concluded after pilot experiments that we needed to allow for as few as 5 due to the issue of participant no-shows. In fact, the fact that larger groups hinder individual contributions is precisely a key mechanism by which group music-making is predicted to enhance bonding relative to speech. We have added clarification to our justification as follows:

> 15-30 participants who are available to come for one hour to a given room during a 3-hour period on a given day are recruited and randomly assigned into three groups of 5-10, each of which is asked to come to the room for a 1-hour slot **(see "Sampling plan" for justification of group sizes of 5-10 balancing theoretical and practical trade-offs).**
>
> …

> Through initial consultation with potential collaborators, we determined the optimal sample size that would allow us to maximise diversity across many sites while allowing experimenters to feasibly recruit relatively large groups of participants was up to 30 participants per site (max 10 per condition across three conditions) for **each of** the sites shown in Fig. 3. Pilot experiments suggested that getting all participants to show up at the agreed location on time was a major unavoidable logistical issue, and that groups of 4 or fewer may not be large enough to test the predictions of the social bonding hypothesis (since singing in small groups "might only produce a small increase… relative to conversation"5). We thus decided to allow for experiments to run if at least 5 participants assembled on time for a given group). Note that while the social bonding hypothesis also predicts that the bonding advantages of singing should increase with even larger sample sizes **("while making music in pairs might only produce a small increase in dyadic bonding relative to conversation, making music in larger, more complex groups of people [dozens or hundreds organized into differentiated sub-groups] should be more effective for collective bonding than language"), and we would have ideally preferred to recruit larger samples per group,** the current experiment is not designed to test this specific prediction of the hypothesis, since such large samples are not feasible to recruit across many sites.

b) l. 201-222. I am also wondering whether the sessions will be filmed and whether one can assess whether singing/reciting/conversation actually took place and whether all participants were involved in the task. I understand that there is no experimenter in the room.

Thanks - we have added the following text in response to this point and to Dr. Hannon's review above:

> **Each experiment will be monitored in real time by the local experimenters via Zoom video, where the instruction video will also be shown using screen share (the experimenters' video and audio will be muted). These videos will not be published, but will be used by the experimenters to monitor compliance and allow them to intervene if subjects misunderstand instructions, in case of an emergency, etc.**

c) l. 190, pre-interaction phase. The instructions say that participants should not interact with each other, but how can one avoid looking/smiling at each other? Perhaps some more specific instructions of what not interacting means may be useful to participants.

We have experimented with different ways of minimising interactions in pilot experiments, but found that the text currently shown in the manuscript and video protocol, with the phrase "Without interacting with the other participants", helps to minimise interaction without overwhelming participants with text instructions, but have now changed the text to red to further emphasize as we have had some issues with following this instruction. We also found that adding a sign on the door of the experiment room helps, and have updated the manuscript with this as follows:

> **To further minimise pre-experiment interaction, a sign with the following text is placed on the experiment room door: "*Welcome to our study. Please enter quietly and do not interact with the other participants until prompted to do so. Please close the door behind you & follow the instructions on the screen.*"**

d) l. 202, pre-selection and length of the songs and other conditions: I think that the authors should be stricter on the length of the songs and that the variability is too high. One can assume that forming a bond takes some time and that it makes a difference whether a group sings for 1 vs. 5 minutes (this represents a fivefold increase in duration!). By the way, there is no mentioning of any duration for the conversation condition. Why not make it 5 min for all three conditions? If a song is shorter, it should be repeated until the time is up.

We have reduced the allowed variability from 1-5 minutes to 2-3 minutes. We found in our piloting that in some societies it was difficult to find a song much longer than 1 minute that many people could easily sing together unaccompanied, but we have to decided to compromise by a) adding karaoke-style instrumental backing tracks, and b) allowing experimenters to choose to repeat the song if needed to reach the target time of 2-3 minutes.

We have also clarified that the conversation condition is always matched to the length of the singing condition:

> **The task time in all conditions will take approximately 2.5 minutes. To achieve this, each site has pre-chosen a song that takes between 2-3 minutes to sing (in sites where an appropriate song could not be found, a shorter song will be repeated for 2-3 minutes), while the lyric recitation will be repeated twice as many times as for singing since lyric recitation is typically twice as fast as singing42. For the conversation condition, a timer will be visible counting down from 2 minutes and 30 seconds. In all conditions, participants will be given 5 minutes to complete both the task and the following survey.**

e) l. 209, conversation topic: I am not sure whether song lyrics are an engaging conversation topic, especially if the authors have such a large variety of songs with different themes on their selected list. Certain songs may generate conversation topics much more easily than others (one needs to interpret the meaning of the lyrics before one can make a statement, which can be quite difficult, and I suspect most people will not know much about the background of songs). Some song lyrics may be somewhat confrontational (anthems?), and to be honest, what can one really say about "Happy Birthday"? ;)

Instead, the authors could choose a relatively neutral conversation topic that can be used across cultures, for example, something related to future travel destinations, how to spend an ideal weekend, their opinion about XY, a newspaper headline etc. I think that the topic of the conversation will lead to emotional involvement and that these feelings may affect the outcome. This is actually also true for the song condition, which is why I suggest stricter selection criteria. One could also offer a picture as a stimulus/trigger for a conversation that is the same in all countries. I would not choose anything that is political or related to religion, but something that can easily lead to an engaging conversation. I understand that the respective song lyrics will be part of the statistical model, but it is important that everybody in a group gets quickly involved in an engaging conversation, otherwise the comparison with singing is not justified. Please also mention in the instructions that people should not talk in pairs, which may automatically happen, especially if the authors are planning larger groups.

Thanks for this excellent suggestion. We had debated this previously and initially decided to try discussing song lyrics to control the semantic content, but after more reflection, piloting, and these reviews (and similar point made by Dr. Göritz below), we are convinced that a more neutral topic not tied to music would be better. We extensively debated what conversation topic would be best, and came to the conclusion that, just as with song choice, it would be better for the conversation topics to also vary across sites rather than being fixed. Among other things, this helps to address the concerns about generalizability across experimental stimuli/tasks raised by Yarkoni's (2022) critique of "The generalizability crisis" We have thus the following text, including a section for choosing a conversation prompt in the new Appendix 3 as follows:

> **Note that we have consciously chosen to allow for variation across sites in the choice of both the song and the conservation prompt in order to maximise generalisability37.**
>
> **…**

Appendix 3: Conversation ice-breaker questions:

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", "identify").
-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)

f) l. 214, Recitation: Please be aware of the fact that some song lyrics may contain meter and rhyme whereas other will not, and some will contain lots of repetition whereas others will not. It may thus be important to check the lyrics for these characteristics and to either control for it or to take it into account in the statistical analysis (not in a pooled random effect). One can surmise that meter/rhyme based lyrics will have larger effects on bonding than free verse. I understand that the authors thought that it may be practical and good to use one song, its lyrics and its topic as a conversation topic across conditions, but at the same time, this may introduce other problems. Why not choose a real poem for the recitation condition? One could agree on a poem with several verses that is typical for a given culture (part of an established canon). One can say that reading a poem in a group is an artificial task, but reading song lyrics (with lots of repetition?) is also an artificial task because song lyrics are usually sung and participants know the song. The use of poems would at least make the results comparable to studies which also used poems and be ecologically more valid.

We appreciate the suggestion, but since the purpose of the recitation condition is not to provide an ecologically valid condition but rather to provide a maximally matched control condition linking the synchronous singing condition with the sequential speech condition, we respectfully prefer to keep this condition as originally proposed (following previous similar designs used by Ozaki et al. 2024 and Bowling et al. 2020). Any issues controlling across songs should apply similarly to the sung and recited versions of the lyrics. We have added the following clarification regarding different metric properties within/across languages:

Acoustic analyses may include variables such as synchrony (including both "self-synchrony" to an isochronous beat and synchrony to other vocalizers79), differences in rhythmic/metric/tonal properties of different languages and of different song lyrics within languages. Such analyses will be complex and are intended to be explored primarily in future publications, so we will not specify detailed analysis plans here. The purpose of highlighting them here is simply to explain the need to record audio for future analysis purposes.

g) l. 263, exclusion criteria: I think that the authors need some sort of evidence that the participants really participated in the activities of the three experimental conditions. Even if singing/speaking will be recorded (by how many microphones?), we do not know whether each individual in a group really performed the task. They may not be honest in the questionnaire. I therefore suggest filming the sessions in addition to recording the voices.

We have added clarification that we will monitor the sessions in real-time via video for compliance and other purposes (see response above).

h) I do not see any explicit mentioning of the debriefing. What information will the participants be given after the experiment?

We have added the following clarifications:

- **3) Demographic variables and debriefings: Additional demographic variables will be collected for exploratory analysis, along with a brief debriefing text.**
- **Projector text: "Follow-Up Questionnaire: Please fill out the remaining page of the survey on your device. Thank you! Feel free to leave whenever you finish, even if the other participants are not done."**
- **Debriefing text (from final page of Qualtrics survey): "The goal of this experiment was to measure whether the average change in social bonding before and after the first singing/speaking/recitation condition from your group was greater than the change in other groups who experienced different singing/speaking/recitation conditions first. Please do not discuss the content of this experiment with other potential experiment participants. If you wish to be alerted when the audio recordings and results of our experiments are published, please provide the email address you would like us to use here (optional - you will not be emailed if you do not provide your address here): _____. We thank you for your time spent taking this survey. Your response has been recorded."**

i) List of songs: I am concerned about the large variability among the chosen songs. The selection criteria could be more stringent in my view. The type of chosen songs varies much across the 50 proposed recruitment sites and languages in terms of length, genre, thematic contents, emotional tone, semantic contents, and appropriate age group. The list of songs in the Appendix contains children's songs, Christmas songs, lullabies, folk songs, anthems, pop songs, songs one cannot identify due to the language, etc. I am wondering how singing a children's song or Happy Birthday feels in a group of (young) adults for several minutes. It could feel a bit ridiculous and inappropriate for a given social situation. There is also a potentially large difference in terms of bonding when singing one's national anthem versus a random pop song. More specifically, anthems may evoke strong feelings in some participants, and if the lyrics are part of a conversation task, they may lead to controversial political discussions in which social cohesion is either significantly heightened (in comparison to a typical pop song) or not experienced at all.

This large variability may also affect the engagement in the other tasks because at the moment all tasks are centered around one specific song (conversation and recitation task). Perhaps the authors aimed at using a wide range of songs to increase the generalizability of their results, but the chosen songs types should not vary so much across sites, especially since there is only one song per site. As expressed above, the length of the songs is also critical and should not show such a large variation in duration. If a song is very short, one could repeat it or add several verses (although singing 5 minutes of e.g. Happy Birthday is also problematic for other reasons and makes the task artificial and may not lead to the desired effects). I think the authors should reflect on these issues and propose more strict selection criteria.

Thanks for these points. As described above, we have reduced the variation from 1-5 minutes to 2-3 minutes, and changed the speech topic from song lyrics to generic conversation, which should resolve some concerns. We agree with the points about song choice issues (indeed, we experienced them personally while piloting "Happy Birthday" and other songs). We have further modified the selection criteria as follows, and as a result modified Table S1 listing the chosen songs. (Thankfully, this is easy to do for a Stage 1 Registered Report, which would not have been the case for a standard non-Registered Report study.):

> **Appendix 2: 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 (though they can read the lyrics while singing). 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.

Analysis plan

a) l. 325, independent variable: Are the authors aggregating across the recitation and conversation conditions? Why? The experiment is clearly designed to have three conditions (singing, recitation and conversation). It is unclear why these conditions are merged.

We have corrected this as follows:

Independent variable: Vocal modality (synchronous singing, synchronous lyric recitation, or sequential conversation)

b) ll. 327-356, dependent variable, prosocial attitude: The number of items representing the DV is rather low in comparison to the number of all other collected background and moderator variables. Is there no short scale on prosocial attitude available that has good psychometric characteristics? The study would profit from using an established measure, rather than a hodgepodge of items.

Thanks for this suggestion. We have added two additional variables ("I feel strong ties to the other participants" and "I identify with the other participants") to create the aggregate dependent variable, with the following rationale:

> **For our confirmatory analyses, we will follow Rennung & Göritz, who previously collated different 17 different self-report measures of bonding/prosociality (perceived similarity, closeness, and liking) used in previous studies and condensed them into 9 variables after removing items with "inadequate discriminatory power, difficulty, and homogeneity"59. From Rennung & Göritz's 9 variables, we excluded the following three variables to minimise redundancy and ensure that the questions could be interspersed with non-bonding-related questions in the questionnaire without making it overly obvious that we intended to measure social bonding:**
>
> **"I have a lot in common with the other participants"**
>
> **"In general, I'm glad to be a member of this group of participants"**
>
> **"I feel affection towards the other participants"**
>
> The final set of 6 variables we used to create our social bonding score are:
>
> 1) "I feel I am on the same team with the other participants"
>
> 2) "I am similar to the other participants''
>
> 3) ''I trust the other participants"
>
> 4) Inclusion of other in the self (IOS): ''How close do you currently feel to all the other participants?"
>
> **5) "I feel strong ties to the other participants"**
>
> **6) "I identify with the other participants"**

c) l. 364, the inclusion of a random effect is good, but as I explained above, the study would benefit from controlling for the length of the condition, the differences in group size, and the type and duration of the selected songs. These points refer to the quality of the manipulation of the experimental conditions and a valid comparison between them. I am not sure whether a pooled random effect is the best way to go ahead. I know that too many random factors can result in a model not converging, but the effect of some factors (particularly those mentioned by Rennung & Göritz, 2016) may be informative. How is "interpersonal dynamics of a given group of participants" assessed?

As discussed above, we have added controls for the length, duration, and type of selected songs and conversation/recitation conditions (all now 2-3 minutes at all sites). As discussed above, pilot experiments suggest it is not feasible to control group size across sites beyond the 5-10 participant per group size chosen. We have deleted "interpersonal dynamics of a given group of participants" since we are not specifically intending to measure this, it was simply an example of one of many possible factors that may vary between experiments.

We have also added the following justification of the random effect modeling:

> **We pool the various potential effects described above into a single random effect rather than explicitly modelling each factor to avoid incorporating too many parameters into the model. Our primary analysis goal is to estimate the fixed effects in three experimental conditions (i.e., singing, recitation, and conversation) under varying factors, including locations, languages, sites, and chosen songs, rather than inferring the magnitude of those factors. Therefore, we consider the decomposition of random effects into multiple factors unnecessary for this confirmatory analysis.**

Pilot data

a) In my opinion, pilot data is only relevant for the exact protocol presented in this registered report. If the experimental setup was very different, as ll. 468-478 suggest, the relevance for this study is limited. The experiment mentioned under 4) seems to use the exact methods as described here and is informative. Again, here (Fig. 4) the authors present three conditions, and I think the statistical analysis should reflect this as well.

We agree that Fig. 4 is the most relevant, and have updated our statistical analysis accordingly to explicitly model all three conditions. We believe it is still worth mentioning

previous pilot experiments to show the evolution of the protocol (especially since there is published pilot data associated that interested readers can access by following our linked references), so have refrained from deleting those brief descriptions. We have also added the following clarification (which also notes new changes introduced in response to this review):

> **Note that the three preliminary rounds (1-3 above) used protocols substantially different from the one proposed here, while the 4th round was almost identical to the current protocol except that the speech condition asked participants to discuss the song's lyrics rather than answer an ice-breaker question, and the dependent variable only averaged the first 4 of what are now 6 items.**

Further data to be collected

l. 510: Moderating variables: I like the list. In case the authors change the instructions of the conditions, 7) and 8) need to be changed. One could also ask how natural the task felt to them, in particular if the authors decide to stick with their choice of reciting song lyrics and talking about song lyrics and songs. It may be informative to ask whether one person acted as a group leader. I am quite sure that if 5 people are asked to sing together, that one person will lead the others, same for the other conditions. It may be informative because synchrony will probably work better if one capable person leads the others during the performance of the task.

Thanks! We have changed questions 7 & 8 from "this song" to "the song on the printed paper" (which is waiting on all chairs in all conditions, as described in the methods) to work with the change in conversation topic from song lyrics to a generic topic. We have also added karaoke-style instrumental accompaniment to facilitate synchronous singing even in the absence of a strong leader. We have also added an exploratory variable about leadership:

> **9) Leadership: "One of our group acted as a leader in the task"**

l. 547, would it not be interesting to ask how much they enjoy singing in general and how often they sing in everyday life? If the authors decide to work with poems, one could ask how much they like reading and literature.

We have added the following variables:

10) Singing enjoyment: “I enjoy singing”
11) Singing frequency: “I sing regularly”

l. 553: Additional data types: I understand that studies comprising many sites are difficult to organize and to lead, and that certain groups may want to collect further data and have different interests, but it is important that the protocol described here is the same across sites and strictly followed, and that participants are not wired at one site and not at the other etc. We should avoid even more variability, particularly if this information is not being present as a factor in any analysis (except in a large pooled random effect…). The current task is short, and therefore I do not see a reason for adding more variability to the set up. It can be easily accomplished in any psychology laboratory on this planet for its own sake.

Great points. As discussed above in response to Dr Hannon, we agree it is worth collecting additional data on cooperative games as suggested, as it can be easily implemented after collecting primary data without affecting the rest of the experiment. However we have removed other additional data types (e.g., EEG) that might introduce the kinds of biases mentioned here.

l. 566, synchrony: See my previous comment about filming.

Implemented (see above), deleting the following accordingly since it is no longer needed now that experimenters are monitoring via video in real-time:

~~“This will also allow us to keep a record of data collection to verify that participants followed instructions despite the experimenter being not physically present in the room.”~~

l. 573, post-experiment conditions. I am confused: “participants will be asked to do the other experimental conditions”? I thought that the experiment is a between-subjects design (see l. 171). I am surprised about the idea to make them do all other conditions after the real experiment to see whether cooperation continues to increase after doing multiple conditions in order. Why not run a within-subjects design in the first place?

Manuela M. Marin

The goal of these post-experiment conditions is not primarily to measure bonding, but rather to collect matched acoustic data for future studies. The additional bonding data is just because we will be doing those conditions anyway. We originally considered doing a within-subjects design but concluded that order effects would make it difficult or impossible to interpret. We have added clarification as follows:

> Post-experiment conditions: **For the purpose of exploratory analysis,** after completing the primary experimental intervention (singing, conversation, or recitation) and survey, participants will be asked to do the other experimental conditions, plus an alternating singing condition (taking turns singing one line at a time). **The primary goal of these conditions is** to enable future acoustic analyses replicating within-participant comparison of the same participant solo singing vs. solo speaking42. Following all these conditions, we will ask them to repeat the social bonding measures again to explore whether bonding continues to increase after doing multiple conditions in order.

Review by Anja Göritz, 26 Sep 2024 08:59

The material to be sung or conversed over or recited is to be a song. This introduces a default bias in favor of singing over speaking. Song could be superior to speech in calling forth prosocial attitudes partly or entirely because of this confound (i.e., being the default) and not because it is causally superior. A clearer test of the hypothesis is possible by using material that either is new or that in any particular site of experimentation is spoken as well as sung at about equal frequency, for example, a (sung) poem that is equally frequently spoken and sung; the first raising procedural issues of training/familiarizing participants with the new material prior to the experiment, the latter raises issues of finding good material.

We have changed the conversation condition from discussing song lyrics to a generic conversation topic not linked to music to avoid this issue (see responses to Dr Marin above).

There was a lack in conceptual clarity in some spots. For example, Table 1: "Contradicts null hypotheses that music is biologically “useless…[c]ompared with language…”3 or “does not directly cause social cohesion”4:" should be expressed more precisely.

Thanks for pointing this out. On further reflection, we have decided to split the main research question into two related ones to be more precise, and updated Table 1 as follows:

Table 1 | Registered Report design planner (simplified overview adapted from https://osf.io/sbmx9; see main text for details)

Question	Hypothesis	Sampling plan (e.g., power analysis)	Analysis Plan	Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis	Interpretation given different outcomes [proposed Stage 2 title wording]	Theory that could be shown wrong by the outcomes
1. Does music **enhance** social bonding?	H1) Synchronous singing **enhances** social bonding relative to a pre-experiment baseline	Maximum feasible sample size: n= 1,710 participants across 57 sites (minimum: 450 participants [150 x 3 conditions] across 30 sites after all exclusions)	Single-blind* multi-site randomized controlled trial. GLMM of social bonding as a function of **time (pre- vs post-experiment; within-subjects)**, with experimental cohort as random effect.	Maximum feasible sample size determined by multi-site recruitment logistics; recommended[1,2] Bayes Factor threshold of 3	if $BF_{10} > 3$: "*Synchronised singing enhances social bonding*"	"*music does not directly cause social cohesion*"[3]
	H0a) Synchronous singing **does not enhance** social bonding relative to a pre-experiment baseline	(same as above)	(same as above)	(same as above)	If $BF_{10} < 1/3$: "*Synchronised singing does not enhance social bonding*"	"*musical behavior is not only associated with, but may causally support, social bonding*"[4]
2. Does music enhance social bonding **more than speech** does?	H2a) Relative to a pre-experiment baseline, synchronous singing enhances social bonding **more than conversation** does	(same as above)	(same as above, except as a function of **modality (singing vs. conversation; between-subjects)**	(same as above)	if $BF_{10} > 3$: "*Synchronised singing enhances social bonding more than conversation does*"	music is biologically *"useless…[c]ompared with language…"*[5]
	H2b) Relative to a pre-experiment baseline, synchronous singing enhances social bonding **more than synchronous recitation** does	(same as above)	(same as above, except as a function of singing vs. **recitation)**	(same as above)	if $BF_{10} > 3$: "*Synchronised singing enhances social bonding more than synchronous recitation does*"	(same as above, f different manifestation of "language")
	H0b) Relative to a pre-experiment baseline, synchronous singing **does not enhance** social bonding **more than conversation or synchronous recitation** does	(same as above)	(same as above, except as a function of singing vs. **conversation and recitation combined)**	(same as above)	If $BF_{10} < 1/3$: "*Synchronised singing does not enhance social bonding more than speaking does*"	music is "*more effective for collective bonding than language*"[4]
If multiple hypotheses are supported, we will combine different types of wording in the Stage 2 title. Possible examples include: "*Synchronised singing enhances social bonding more than conversation or synchronised recitation does*" or "*Synchronous singing enhances social bonding, but not more than speaking does*". If all $1/3 \leq BF_{10} \leq 3$, we propose *"Inconclusive evidence for effects of synchronous singing on social bonding"*						

**Following ref.[6], we classify this experiment as "blinded" because the experimenters will be "not present during the manipulation and measurement of outcome variables".*

Lines 78-81: The quote does not pertain to the statement in 78 and 79. What means "directly causing", and is signaling not a way of causation?

We agree that signaling can potentially be a causal mechanism for bonding, but we believe that in this case Mehr et al. are arguing against such a causal relationship. We have changed the statement in (what were previously lines) 78 and 79 so the quote now pertains to the statement. The sentences now read:

> **Nevertheless, some remain skeptical that music specifically causes social bonding 3,6,28–31 (but cf. 32). For example, Mehr et al. wrote:**
>
> **"*music does not directly cause social cohesion: rather, it signals existing social cohesion that was obtained by other means*".3 [emphasis added]**

Lines 116 & 128: The cited references could not be accessed without installing a plug-in, which I did not do. This barrier to peer review should be removed.

The plug-in links are not required to access the references - you just need to scroll down manually to the reference section to the corresponding numbers (e.g., reference #21, reference #46).

Having merely self-reported prosocial attitude as a dependent measure is methodologically weak. A behavioral measure of cooperation/cohesion should be added, for an example, see https://doi.org/10.1371/journal.pone.0136750. Although for some sites of experimentation, something along these lines is planned, it would be better to do it in all sites, also working toward across-site standardization. Pretesting and calibration of the implementation of this bevarioral dependent measure should ensure that ceiling/bottom effects are unlikely.

We have implemented this suggestion at all sites (see response to Dr Hannon above).

Line 325: Given that the independent variable is dichotomous (singing vs. speaking), but that there are three independent groups the "matching" remains unclear. Is the singing group compared to the collapsed recital and conversation groups? What if singing is inferior in calling forth prosociality to one of the speaking groups but not the other?

We have changed the description of independent variable and hypotheses to specifically distinguish between the synchronous recitation and sequential speaking conditions (see responses to Dr. Marin on these points above and Table 1 copied above).

Line 327: "Cooperation" is a long way from what is assessed in this experiment. The naming should be more cautious and proximate to what is assessed. For example, "self-reported attitudinal prosociality" is a better choice.

Good point, similar to Dr. Marin's. We have changed the description of the dependent variable here to "Social bonding (Self-reported attitudinal prosociality)". We have also changed "cooperation" to "social bonding" in the title and throughout the manuscript.

Lines 216-218: "participants will be asked to repeat the lyric recitation twice to ensure it takes a similar amount of time as the other conditions": There is a dilemma with regard to interpretability/internal validity that should be addressed: either the time is the same but number of repetitions are unequal, or the number of repetitions are the same but time is unequal. Best is to test either horn.

We do not understand the meaning of "Best is to test either horn", but we have clarified this dilemma/decision as follows:

> **We acknowledge that this introduces a different confound, but given the choice between controlling the time or the number of repetitions, we felt it more essential to control the overall interaction time to match both singing and conversation conditions (since in any case the content of the conversation condition is also different from both singing and recitation conditions and different songs also often have varying degrees of internal repetition).**

Since the content/manner of the conversation is not scripted and thus uncontrolled, the effect of conversation to call forth prosociality might highly depend on the spontaneously unfolding of content/manner of the conversation. For example, if humor comes up or particular insights come up this might be especially (un)evocative of prosociality.

This is true. We have decided that this is more of a feature than a bug, so have decided to further enhance variability of conversation prompts to maximise generability, as follows:

> **Note that we have consciously chosen to allow for variation across sites in the choice of both the song and the conservation prompt in order to maximise generalizability37.**

We have also added this explicitly as one of the factors modeled by our random effect as follows:

> We model each cohort of 5-10 participants using a random effect to account for a number of variables that may (co)vary between groups, including (but not limited to):
> -differences in group size due to no-shows
> -language spoken
> -song chosen (e.g., musical/lyrical/symbolic content, amount of repetition)
> **-content of conversation prompt and spontaneous discussion in a group**
> -cultural values (e.g., individuality/collectivity, norms about group singing/speaking)
> -experimenter effects (e.g., physical set-up of the experiment room, method of participant recruitment)
> -time of day
> -geographical location

-etc.

I advise to not collect demographics 7-11 because they give away the research question or point participants to the emphasis on music. This being an experiment any demographics and auxiliary variables can and should be restricted to the minimum. Some of the other demographics/auxiliary items appear dispensable, as well.

We had originally intended to collect all demographic variables after the experimental phase for this reason, but then decided to move them before the experiment to avoid conditioning on post-treatment variables. However, upon further reflection after seeing your suggestion, we are convinced that collecting all demographic variables post-experiment is most appropriate after all. Since all demographic variables will only be used for exploratory analyses anyway, this should maximise the strength of conclusions that could be made from our main confirmatory analyses. We have changed the manuscript to reflect this as follows:

> Note that all demographic variables will be collected **after collecting all data used for confirmatory analyses, in order to minimise potential expectancy effects28 and any other potential unintended effects on participants during the main experiment phase.** ~~**during the initial baseline condition prior to the experimental task to avoid conditioning on post-treatment variables72.**~~