Author Responses to Reviewer Comments

July 22nd, 2022

General Responses from the Authors

We thank both reviewers and the editor for many of their comments concerning the content and wording of our submitted manuscript. After many revisions, we believe that we have attended to every comment either by modifying this the text and noting how we have attended to this or in our review comments below. We have annexed each of our comments below your original comments beginning with a \*\* to denote our collective response.

The major points we point out concerning our revisions deal with how we address the issue of “replication” in the study. We agree that our original presentation of our data and analysis lead to misunderstanding concerning replication and have since re-organized how we present our five hypotheses and corresponding analyses. Each hypothesis is now more clearly mapped to each analysis and given more motivating context. Our analyses now also conform to APA7 standards in terms of the statistics being reported in the text with relevant p-values and have re-done several of our figures to make our results clearer and in-line with Musicae Scientiae house style.

Reviewer(s)' Comments to Author

Reviewer: 1

Comments to Author

This paper reports a replication study pertaining to lyric effects on emotion ratings to music. In this case, the replication was specific to the vastly known research by Ali and Peynircioglu (2006). This topic is important and impactful, in that lyrics have received less attention compared to musical phrases and surface features in the cognition and psychology of emotion in music, thus numerous questions remain. I commend and recognize the work that has gone into the research and its write- up. While this is a replication study and thus does not present novel knowledge, the importance in replication/generalization of methods and results is validated.

However, three are limitations to the study; some relate to the communication of ideas while others, and most importantly, refer to the conceptual framing of the work, which later informs the method. I hope my review will assist in developing the program of research, including for publication.

Major point:

The reference to Ali and Peynircioglu (2006) is well-grounded. The design, and the analysis of data is rigorous and well presented. However, how can we call this a replication study, when the stimuli and culture presented to is not being replicated? There are serious limitations to this, which are not explained in the manuscript. The specific effect from the psychophysical and/or cultural cues are not going to be re-tested, and hence this becomes a highly questionable replication study. To simply explain it as a replication, when little is being replicated, is not very elegant, and requires explanation. By investigating responses made by Western listeners to Indian ragās does not imply cross-cultural validity. Nevertheless, a simple solution to this would be to remove this concept of ‘replication’ and just highlight that this is a study to investigate the effects of lyrics. There is absolutely nothing wrong with this, but it needs to be made clear from the onset. The reference to replication makes a reader assume some expected systematic comparisons. If the interest is replication and reliability, further research is required which would involve (1) identical stimuli and, (2) similar participants. Thus, in my view, I think the article would be much improved if it did more comprehensive work on this, or simply, omit the link to replication and reliability altogether.

Furthermore, simply because the 2006 is well cited is not reason enough to run a replication/reliability study. If anything, consider replicating the research because of its limitations, which they very well explain. As such, please consider rewording your title: Generalizing is a big claim, and I believe it would require significant more work to generalize the effect of lyrics on emotion ratings after conducting one study.  
There is important research missing. Please consider including and discussing the such works which specifically investigate lyrics and emotions (for instance, Barradas & Sakka, 2021 – Psychology of music; Susino & Schubert 2019 - Empirical musicology review and Susino & Schubert 2020 – PLOS One; Brattico et al, 2011 - Frontiers).  
There is no information on the pilot study. This needs to be clearly explained, because a large part of the methodology is based around the pilot. Please include all details about this pilot.

\*\* Thank you for your kind words about many aspects of our manuscript. We will address the major points you bring up here and the smaller points below, in turn. Your first major concern centers around our use of the word replication. As we are not using the exact same materials and design, it would not qualify as what we would call a direct replication and we agree with this concern. Our use of the word replication follows Nosek and Errington’s (2020) definition in that “Replication is a study for which any outcome would be considered diagnostic evidence about a claim from prior research. The definition reduces the emphasis on operational characteristics of the study and increases on the interpretation of possible outcomes”. As we were originally interested in the same factors and levels of the original study (presence of lyrics [present-not], emotion[happy-sad-calm-angry], and gender[male-female]) in a design where the outcomes would tell us about the extent to which each of these contributed to ratings in a within-subjects design, we feel that our results (showing significant effects but in different directions) allow us to make a diagnostic claim about previous evidence (that presented in Ali and Peynircioglu 2006). In order to make this assumption direct, we have now added a direct reference and footnote to the paper that provides this specific definition of replication. We respectfully disagree with the assertion that for a study to be a replication it needs both (1) identical stimuli and (2) similar participants. As noted in the Nosek and Errington paper, this criteria would render almost any study unreplicable as it would demand a single stimuli set for any effect reported in the literature. This also is what Yarkonki (2020) discusses as the “Fixed Effects Fallacy” in which empirical researchers believe that they are able to generalize an effect (in this case, effects of lyrics and emotions put forward by Ali and Peynircioglu 2006) based on a fixed set of stimuli when in reality to make any claim from stimuli to effect, this needs to be modeled as if it could be other stimuli that preserve the same properties. A similar argument regarding how similar participants need to be follows the same logic.

Regarding your second point, we very much agree that motivating this study (and hopefully others in the future!) is better done by highlighting some of the limitations of the original. As you note that we have explained this well, we have only added an additional sentence to the opening paragraph that uses the citation introduction as a means to discuss the limitations of the study.

We also agree that the title over states our findings and discussion; thus, we have changed the title to be “Lyrics and Melodies: Do Both Affect Emotions Equally?”.

Considering your last major points, we have added all these citations where appropriate (the Brattico et. al (2011) is already cited in the submitted version of the manuscript). We did not include any information about our pilot study as we were initially constrained for space in the initial submission. We have added a paragraph giving a cursory description of relevant information about our pilot study.

Minor points:

Lines 5: Does it really matter?

\*\* We assume that this is referring to Page 1 of the manuscript in our reference to citation count. For many, citations are used as a proxy (albeit not that reliable) for impact (at least within the UK), which we wanted to use as a narrative device to capture the reader’s attention and contextualise this specific paper within the context of every other paper. We have re-written so the attention focuses more on the limitations of the study.

Line 35: Why the circumplex (Russell 80) citation? Seems like a throw away.

\*\* We do not see this as a “throw away”, because the original paper decided to select their stimuli based on the two axes of Russel’s model (valence, arousal), leading to the selection of the four categories used in our experiment. We would like to keep this reference for readers who have not read Ali and Peynircioglu (2006).

Line 40: Please explain how this is not contradictory to replication? A simple delete might suffice.

\*\* Regarding this comment, please refer to our response on the main concerns above. To briefly summarize, we do not think that exact stimuli are needed for replications, as that would always tie any reported effect in the literature to the exact stimuli set used in an experiment. We contacted the original authors and asked for the stimuli, and were told that the stimuli were no longer available. Thus, in cases like these, one could never try to address a previous finding if the exact stimuli were required for any attempt.

Line 47: Please explain what replication crisis is.

\*\* We have added a sentence to explain this concept briefly in the revised manuscript.

H1: “In our replication analysis we predicted we would directly replicate the analysis” ... this is hard to understand. Please consider rephrasing.

\*\* We have reworded this to say “In our replication analysis, we predicted we would find a similar pattern of results as reported in Ali and [Peynircioğlu](https://paperpile.com/c/P5SZxy/LOoz) (2006).”

Line 28: 74 participants? But before it was 108? And: participants were, not ‘are’.

\*\* The 74 refers to the number of participants from the department of psychology. Later that sentence, we list the number from the school of music (34). (34+74 = 108). We have changed “are” to “were” to keep the tense consistent across the paragraph.

Page 9, line 37: But before it was mentioned they did not explicitly mention the locus of emotion, so how is it now ‘felt’ emotion locus?

\*\* The original paper of Ali & Peynircioglu (2006) did not mention the specific locus of emotion. In our study, we clarify this both in the Methods section (paragraph 2 of Design) and the Discussion (Paragraph 3 of the heading Replication) .

Reviewer: 2

Comments to Author

The current manuscript is a replication attempt of Experiment 1 of Ali & Peynircioglu (2006), which investigated the effects of lyrics on emotion across happy, sad, calm, and angry music. In the original paper, Ali & Peynircioglu (2006) found a main effect of emotion: higher ratings for happy music than sad/calm/angry music. They also found a main effect of lyrics: music without lyrics was rated as more intense than music with lyrics in general. There was also an interaction between the intended emotion of the music and whether or not there were lyrics: music with lyrics led to lower emotion ratings for happy and calm music, but higher emotion ratings in sad and angry music. The current study showed different effects: there was a main effect of emotion, but calm music was rated as highest, with differences between conditions. There was a main effect of lyrics but it was reversed – music with lyrics was rated as higher than music without lyrics. And there was an interaction, but the direction of this interaction was not outlined in the paper, and it is not possible to see with the graphs presented.

The paper provides a nice outline of the replication crisis and why it’s valuable to replicate findings. The study also had a large sample size, and the statistical analyses run were appropriate. However, it is missing some theoretical background to the importance of the experiment, and the authors made some important methodological changes which could have affected the results, and should be discussed in more detail. Further, not all statistical results and the direction of these results are reported, which make it somewhat difficult to fully understand.

Details outlined below:

Major Points:

The introduction discusses the importance of replication in psychology, which is great, but can you also provide more of a theoretical background to the current experiment?

Why did you decide to replicate Experiment 1 of this paper in particular (why not other papers, why not other experiments in this paper), and why is it important (this is mentioned briefly, but I think needs to be outlined in more detail).

What is the theoretical reason that this study was run in the first place? It needs to be linked into the literature again, for the moment the theoretical reasoning is not clear.

\*\* Thank you for your comments! We list out what we have done to attend to each of these concerns here. First, in order to make clear why we were interested in the first experiment explicitly, we have added sentences at the end of the last paragraph of our introduction (Beginning p. 2 “In this paper we report...) noting that the first study was the motivation for the three other studies reported in the paper and that they last two papers reported in the study deal with visual associations on emotion, which was beyond the scope of our initial study. Based on comments from R1, we also have linked in the literature that also engages with the findings put forward by Ali and Peynircigolu (2006), which we believe better links this to the literature/theory at large. Please note that we use language here that is a bit more broad in order to attend to R1’s comments regarding using the term “replication” for our study.

Pg. 5, methods: The stimuli construction has introduced some big differences compared to the original study which I’m concerned about. First, the melodies in the replication seem to be made using a MIDI keyboard, which would arguably remove much of the emotional content (expression, dynamics, etc). The original music appears to be full musical excerpts, presumably containing multiple instruments, expression, dynamics etc. In the original paper, they say “the melodies were taken from original source recordings and the vocalists recorded the lyrics to go with these melodies”.

It would be interesting to see the direction of the interaction between lyrics and emotion here (this missing information is outlined in a point below) - is it in the same or different direction as the original paper? This could be related to the additional vocal cues perhaps. This needs to be outlined in more detail in the discussion - these stimuli changes seems a more plausible reason for the difference in results. Perhaps in MIDI music calm music is more easily communicated than happy music (for example).

\*\* Concerning the use of MIDI, this might be a confusion in that MIDI refers to the way the musical information is encoded and nothing about the quality of the timbre. To clarify this in the manuscript, we have added a sentence to note that the sounds were made with a professional grade instrument played by a professional pianist. We have also added information about the direction of the interaction and discussed it in relation to the original paper but do this in a point below (as you note in other comments). Regarding the post-hoc theorizing of the direction of this interaction, while we could attempt to guess why this interaction might be the case, we feel this would qualify as TARKING (theorizing after data collection), which runs counter to our later caution to not over-interpret results of our empirical findings without a clear theory (since many theories presumably can generate the same pattern of data). Presumably many data generating processes could have contributed to our findings, and we do not want to speculate what else this could have been.

Second, adding a voice cue “no” to the no-lyrics condition adds a new element that was not in the initial study. The choice of this syllable does not appear to be neutral (as stated), as “no” has a negative connotation in English. Adding a vocal cue might also introduce confounding levels of emotion (were the syllables sung with the intended emotion?). The addition of a sung word also makes the assumption that the reported results in Ali & Peynircioglu were based on the WORDS of the lyrics, rather than the vocalisation itself. This is not necessarily the case. I’m wondering if the reversal of the lyrics effect might be related to the vocalisations in the no lyrics condition, which could somehow be more distracting.

\*\* We agree that there could be some sort of semantic effect of the syllable; however, this was the syllable that was suggested to the team to provide the clearest open sounding vowel. We have added a sentence in the Discussion to suggest that future research could also consider this as a potential confound. (Paragraph starting “In considering the difference in findings…)

Pg. 4, Participants. Please report the musical training of these participants. Since 34 were recruited from the school of music, these are likely to be highly trained participants, unlike the participants in Ali & Peynircioglu. Looking only at the 74 comparable participants, do the results change?

\*\* The question regarding musical training is addressed in the Extension analysis. In this analysis, we report similar findings in regard to the effect of lyrics and emotions, with additional significant contributions related to the emotional engagement but not musical training GMSI subscore. We would expect that there would be groupwise differences were we to dichotomize based on the selection biases of students who were in the school of music, but resist dichotomizing and instead measuring these factor continuously based on arguments put forward in the original GMSI paper (Mullensiefen, et. al, 2014) and arguments put forward by statisticians (Harrell, 2015; Royston, Altman, Sauerbrei, 2005 (<https://doi.org/10.1002/sim.2331>) but are perhaps beyond the scope of including in this paper.

Pg.8, lines 5-22 – it is unclear how the linear mixed-effects model was run. It says that “emotions rating” is being predicted – but which emotion rating is this? For each subscale? Or collapsed across all? And then how do the conditions fit in? And why is there no happy condition when you have the other three conditions? Same with the interactions. Please make this clearer.

\*\* We have clarified this by adding the following sentence and given an example in the text. In Hypothesis 3 paragraph, it now reads “A main effect of combined emotion category such that music in the positive category would be rated higher than music in the negative category, a main effect of lyrics such that melodies without lyrics would be rated higher than melodies with lyrics, and a potential interaction between emotion and lyrics.” As for “missing” happy conditions and interaction, these are in the table and are a result of coding the model with “happy” as the baseline, with each other coefficient noting how the estimate changes from baseline.

Minor Points:

Pg. 3, line 31-35: “musical training might explain variation in responses”. Do you mean to emotional responses in particular?

\*\* This has been changed.

Pg. 3, line 49-54: The sentence “this hypothesis included the main effect of lyrics and emotion...” is difficult to understand. Please rephrase or make two sentences.

\*\* We have split this into two sentences.

Pg. 5, methods: why did you choose 35 seconds and not 20 like in the original study?

\*\* We added the sentence “We changed the length of clips from the original 20 seconds to 35 sot hat we could include two phrases of lyrics in all songs presented to participants” to show this was done for practical reasons.

Pg. 5, methods: Participants heard BOTH the lyrics and no-lyrics version of each song. Perhaps this also changed the second rating? (in the original study they only heard one version of each).

\*\* We considered this, but believe we have addressed this concern when in the first paragraph of Design, we note that “ Participants heard both blocks, with the order of blocks randomized and stimuli randomized within each block. Songs were distributed so that participants did not hear the same song both with and without lyrics in the same block”.

Pg. 5/6, make it clear that you only analyse the rating from the intended emotion. Did you look at the ratings for the other emotions? Ali & Peynircioglu verified that the intended emotion was rated the highest. There might also be some interesting data to see overlap (e.g., are “positive” emotions confounded most often, or “negative” emotions? )

\*\* We agree that this analysis is interesting and have code to produce these analyses in the supplementary materials and is noted in the paper. We did look at that, but from this comment, we were not sure if you were requesting that we add this analysis to the paper so we did not (also since this is under “minor points”). If this is not the case, please let us know explicitly. We have decided not to include it for the R+R here as this was not listed in the major considerations above.

Pg. 6, line 38: “we report only main effect of gender and lyrics, with no significant main effect of gender or any interactions”. Do you mean emotion for the first “gender”?

\*\* Yes, thank you so much for pointing that out. This has been changed.

Pg. 7, top: Please write out the results and exact p-values/effect sizes from the Tukey HSD tests, and show how you control for multiple comparisons. The significant pairwise difference between the lyrics condition should also be written out, as well as the direction of the difference.

\*\* The Tukey HSD test compares all possible pairwise comparisons and divides an alpha of .05 by the number of all possible tests (Cohen, B. (1996). *Explaining psychological statistics.* Thomson Brooks/Cole Publishing Co., 4th edition, p. 412 ). Our choice of using a Tukey HSD reflects how we decided to control for Type I error/multiple comparisons. These differences are also shown visually in Figure 2.

Pg. 7, group analysis: explain the direction of the main effect of lyrics, and what the interaction means? (Assuming this is an interaction between emotion and lyrics? Please specify).

\*\* We have added language to note the means in the main effects. As we have provided four means (lyrics/none + positive/not) the pattern of interaction can be reconstructed for readers. We have added a sentence describing it verbally.

Pg. 7, line 40-42 – It wasn’t mentioned earlier that there were musical and emotional subscales to the Gold-MSI – please mention so that this analysis doesn’t come as a surprise.

\*\* We have added language to the Methods that specify that the Gold-MSI has several subscales. This is found in the second last paragraph of the section noted as “Design”.

Pg. 9, 17-22 – “This outcome is reversed from the original study and suggested that under the current context, the initial findings were not robust and did not generalise”. Make it clearer the differences in stimuli that contribute to the current context.

\*\* This point is now clearly and explicitly stated in our discussion. See paragraph beginning “In considering the difference in findings,”)

Pg. 9, lines 33- 36: current study says that you ask about perceived rather than felt emotions (as they did in Ali & Peynircioglu). It’s unclear from the original paper whether they were assessing felt emotions. They report “the participants rated their EMOTIONAL judgments on response sheets. Further, the options were broader, e.g., in one category they were happy, joyful, exciting, and/or festive; or sad, depressing, and/or melancholy. Perhaps this wider range of emotions had an effect compared to more categorical happy/sad/ etc. in the current replication?

\*\* We agree this might have contributed, but also note that Ali and Peynircioglu did not specifically say if they requested felt or perceived emotion. This is currently addressed in our discussion where we state “In addition to using stimuli based on the design, we further prompted participants to explicitly respond based on what emotions they perceived, rather than felt. This differs from the original paper, which only asked for ratings and did not make a clear distinction between felt and perceived emotion--that possess different mechanisms and influences on emotion ratings”. We agree that having more descriptors (a wider range of emotions) might have had an effect here, but we also would argue that this wider range of emotions would be more correlated with one another than the other groupings. This is also why we explicitly have declared our assumptions in using Russell’s Circumplex model.

Pg. 10, line 20-23. You suggest a lack of theoretical support for the initial hypothesis, but do not add any theoretical support for the current replication or hypothesis.

\*\* This is true. Our primary concern in this paper was the stability and robustness of the initial findings put forward, not the data generating process that led to the process. Many processes could have led to either our data or the original dataset and we consider that discussion beyond the scope of our replication study, and would not be possible without specific experimental interventions.

Pg. 11, 19-27: you could test the statistical artefact idea by running the same analysis with only musical training OR emotional engagement. If this is true, then musical training would then become significant.

\*\* We have run both of these analyses and have updated the analysis and discussion accordingly to state that running this analysis without the emotional factor does lead to a significant result of training. We add a sentence highlighting this and to note that a common factor could be driving this finding.

Pg. 11, line 33-38: why would the current findings corroborate claims encouraging measuring aspects of music perception differently from music production? This is not evident. Is this meant to be in link with musical training (which is both perception and production, and not based on a task).?

\*\* We agree. We have changed the sentence to read “We believe our findings corroborated claims encouraging measuring aspects of musical engagement with scales and tools that make conceptual distinctions between the various ways in which an individual can engage with music”. This makes it clear that researchers must be careful as to think about what mechanism might drive these effects and how to best measure them.

- Was there any analysis of familiarity ratings?

\*\* Yes, this was done and can be found in the supplementary materials. We have added a sentence in the results section noting this was done, that it is supplementary materials, and had no effect.

Typos: Pg 1, line 22: Should be Fiveash, not Fiveas.

\*\* Thank you, this has been corrected