# Lyrics and melodies: Do both affect emotions equally?

Cited over 240 times, the article “Songs and emotions: Are lyrics and melodies equal partners?” (Ali & Peynircioğlu, 2006) reported an investigation of whether the presence of lyrics in a piece of music affected listeners’ perceptions of the emotion conveyed by the music. A recent bibliometric analysis of the music psychology literature shows that this article is in the top 15% of papers cited in music psychology (Anglada-Tort & Sanfillipo, 2019).[[1]](#footnote-1) Ali and Peynircioğlu’s findings continue to enjoy attention, with contemporary researchers drawing on them in more recently published work (Akkermans et al., 2019; Dahary et al., 2020). The knowledge that lyrics can affect listeners’ perceptions of the emotion conveyed by music establishes the expectations of researchers using similar designs (Barradas & Sakka, 2021; Fiveash & Luck, 2016; Pereira et al., 2011; Susino & Schubert, 2019; 2020). Given the attention this article has received, we expanded on Ali and Peynircioğlu’s line of research with the intention of addressing several of the limitations of the original study and providing further evidence of the factors affecting listeners’ perceptions of emotion in music.

In this paper we report findings from a replication experiment[[2]](#footnote-2) using a design similar to that reported by Ali and Peynircioğlu (2006; Experiment 1), but with a larger sample size and pre-registered analyses to investigate the relationship between the presence of lyrics and ratings representing listeners’ emotional responses to stimuli. In our study we replicated Experiment 1 only because, of the three subsequent studies it motivated, two addressed associations between visual stimuli and emotion perception, topics beyond the scope of our initial investigation. On the basis of the literature published since 2006 (Akkermans et al., 2019; Dahary et al., 2020; Weth et al., 2015), we included covariates that might explain the discrepancies in the ratings reported by Ali and Peynircioğlu (2006) and the ratings awarded by the participants in our study.

**Replication in music psychology**

As a subset of psychology, music psychology is not immune to what is referred to as the “replication crisis” (Hutson, 2018; Ioannidis, 2005; 2016; Open Science Collaboration, 2015; Peng, 2011), that is, that it is difficult or even impossible to reproduce many findings published in the scientific literature. Frieler et al. (2013) argue that the findings of research in music psychology might be even more unstable than other sub-specialties of psychology because music psychology is “low-gain/low-cost science” (p. 271); scarce resources produce smaller samples and, in addition, music-psychological theories are relatively harmless. Thus, research in music psychology may tend to value novelty, creativity, and originality over validity and reliability. In recent years, however, more replications of studies in music psychology have been published, with topics ranging from psychometrics (Baker et al., 2018; Chamorro-Premuzic et al., 2009; Lin et al., 2019) to physiological responses to music (Bullack et al., 2018), and listeners’ evaluations of piano performance (Wolf et al., 2018).

Ali and Peynircioğlu’s (2006) article has served as an important reference point for more recent articles on listeners’ perception of emotion in music (Brattico et al., 2011; Tsai et al., 2014). In Experiment 1 Ali and Peynircioğlu examined the ratings of 32 participants representing their perception of the emotion conveyed by musical excerpts in four categories, with and without lyrics. The four emotion categories were happy, sad, calm, and angry (Russell, 1980). In addition to these within-subject variables, the authors also included gender as a between-subject variable resulting in a 4 x 2 x 2 mixed-model ANOVA. Full details of their methods can be found in the original article (Ali & Peynircioğlu, 2006).

[ FIGURE 1]

The authors report a significant main effect of emotion category. Post-hoc comparisons showed that this was attributable to a difference only between ratings of musical excerpts in the happy category and the other three categories, as depicted in Figure 1A. They also report a main effect of lyrics such that melodies without lyrics received significantly higher ratings than those with lyrics (Figure 1B). There was no effect of gender, *F*(1, 30)=0.30, MSE=3.81, *p* > .10, not shown in Figure 1, but there was a significant interaction between gender and the presence of lyrics such that women gave lower ratings to melodies without lyrics than men, *F*(1, 30)=4.57, MSE=1.13, *p* < .05, not shown in Figure 1). The authors further conducted a separate analysis in which the happy and calm categories were combined in a single positive category, and the sad and angry categories were combined in a single negative category. Regardless of the other variables, music in the positive category was rated higher than music in the negative category (Figure 1C).

Although the aim of our research was to replicate and extend Ali and Peynircioğlu’s (2006) study, direct replication was not possible as we were unable to obtain the materials they used. We therefore created our own stimuli. Unlike the original study, we explicitly asked participants to rate their perceived emotions to disambiguate the potential confound of felt versus perceived emotions (Gabrielsson, 2001; Juslin, & Västfjäll, 2008).

Second, given the literature on music and emotion perception suggesting that variation in responses might be attributable to individual differences, including extent and nature of musical training (Akkermans et al., 2019; Castro & Lima, 2014; Dahary et al., 2020; Karreman et al., 2017; Taruffi et al., 2017), we used the Goldsmiths Musical Sophistication Index (Gold-MSI; Müllensiefen et al., 2014) as a robust, continuous measure of musical sophistication. This literature motivated our decision to include only two of its subscales in our analysis, Musical Training and Sophisticated Emotional Engagement with Music (henceforth “emotions”), although participants completed all of them.

We predicted that there would be (1) a main effect of emotion category such that music in the happy category would be rated higher than music in the other three categories, 2) a main effect of lyrics such that melodies without lyrics would be rated higher than melodies with lyrics, and 3) an interaction effect such that women would give lower ratings to melodies without lyrics than men.[[3]](#footnote-3) We also predicted 4) 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. Based on the literature on individual differences in music and emotion perception, we hypothesized that 5) scores on the musical training and emotions subscales of the Gold-MSI would predict ratings of emotion category such that participants who scored higher on the Gold-MSI scales would give higher ratings for each emotion category. The analysis plan followed our pre-registration [REDACTED FOR PEER REVIEW] on AsPredicted.org, which can be found in this project’s Open Science Framework (OSF) repository.[[4]](#footnote-4)

# Method

**Pilot study**

All hypotheses, design, and analyses were pre-registered on AsPredicted.org and all materials are hosted in the OSF repository. The results of a pilot study ([REDACTED], 2019) confirmed that participants (*n*=16) understood the task and were not overly familiar with any of the musical excerpts we had chosen as stimuli, and that each excerpt evoked its intended emotion.

## Participants

One hundred and eight students were recruited from the Department of Psychology (*n*=74) and the School of Music (*n*=34) at a university in the Southeastern United States. To be included in the study, participants had to be native speakers of American English, with normal hearing and vision; four students who did not meet these criteria were excluded, resulting in a final sample of 104 participants, 62 women and 42 men with a mean age of 19.2 (*SD*=1.37, range 18-26). Their average mean scores on the musical training and emotions subscales of the Gold-MSI were 24.6 (*SD*=13.3) and 35.7 (*SD*=5.07) respectively.

**Materials**

Because we were unable to obtain the stimuli used by Ali and Peynircioğlu (2006) we created a new set of stimuli based on the original materials. We selected songs with relatively low playtimes in Spotify (see Supplemental Material) to ensure that they would be unfamiliar to participants. Each of the songs, which were in popular and classical genres, was intended to convey one of the four categories of emotion corresponding to the happy, sad, calm, and angry quadrants of Russell’s (1980) circumplex theory of emotion. While Ali and Peynircioğlu used 20-second clips, we recorded 35-second clips from the first verse or the refrain of the song to include two phrases of the lyrics of each song. In a second change to Ali and Peynircioğlu’s method, our recordings were made by a professional pianist on a MIDI keyboard, using the timbre of a professional-grade grand piano, accompanying a singer who either sang the lyrics or vocalized on neutral syllables “no (/nō/)” so that each excerpt existed in two versions, one with and one without lyrics. This ensured that the melodies of the excerpts in the lyrics and no-lyrics conditions were consistent. The 32 recordings were then divided into two blocks, with equal numbers of lyrics and no-lyrics excerpts in each block.

## Procedure

Participants undertook a listening task in which they heard both blocks of stimuli, with the order of blocks randomized and stimuli randomized within each block. Excerpts were distributed so that participants did not hear the same excerpt both with and without lyrics in the same block. After listening to each excerpt, participants were prompted with the following statement: “Please rate the extent to which you agree the music you just heard could be described with the following four emotions. After rating the four emotions, please rate the extent that you are familiar with the music you just heard.” Participants were asked to give their ratings on a 9-point Likert-type scale with five items, the first four corresponding to the degree to which they perceived each excerpt as happy, calm, sad, and angry, and the fifth rating their degree of familiarity with the excerpt they had just heard. When they had undertaken the listening task, participants provided demographic information and completed the Gold-MSI questionnaire (Müllensiefen et al., 2014). Both task and questionnaire were administered in a sound-attenuating environment. The excerpts were presented in jsPsych (de Leeuw, 2015) and participants heard them through Sennheiser HD 229 over-the-ear headphones connected to a PC running Windows 10 at a volume judged subjectively by the research team to be comfortable.

# Results

To investigate our first three predictions, we conducted a 4 x 2 x 2 mixed ANOVA using the ez package (Lawrence, 2016) in the R programming language (R Core Team, 2020). The Greenhouse-Geisser correction was applied to examine the influence of three independent variables (emotions, lyrics, gender) on ratings of emotion. Our emotions condition had four levels (happy, calm, sad, angry), our lyrics condition had two levels (lyrics, no-lyrics), and our between-subjects gender variable had two levels (female, male). The ANOVA was computed with Type III sum-of-squares. Main effects were explored using a Tukey HSD test implemented in the multcomp package in R (Hothorn et al., 2014).

Prediction (1) was upheld insofar as there was a significant main effect of emotion, *F(*3, 264.30)=28.24, *p*<.001, η²G =.11,[[5]](#footnote-5) although excerpts intended to convey happiness were not rated higher for happiness (*M*=5.10, *SD*=2.71) than music in the other three categories (sad: *M*=6.06, SD=2.52; calm: *M*=6.63, *SD*=1.98; angry: *M*=5.56, SD=2.98). Other than the difference between ratings for excerpts intended to convey sadness and calmness (*p*=.37), which was non-significant, all differences were significant (calm vs. happy, *t*=9.00 *p*<.001; angry vs. happy, *t*=3.98, *p*<.001; sad vs. happy, *t*=7.13, *p*<.001; angry vs. calm, *t*=-5.01, *p*<.001; sad vs. angry, *t*=3.14, *p*<.01). Prediction (2) was also upheld insofar as there was a main effect of lyrics, *F*(1, 103)=203.46, *p* <.001, η²G=.10, although excerpts without lyrics did not receive higher ratings (*M*=5.34, *SD*=2.64) than excerpts with lyrics (*M*=6.34, *SD*=2.52). There was no main effect of gender, and Prediction (3), that we would find an interaction between gender and lyrics such that women would award lower ratings to no-lyrics excerpts than men, was not upheld.

To test for a main effect of combined emotion category such that excerpts in the positive category would receive higher ratings than excerpts in the negative category (Prediction 4), and a potential interaction between emotion and lyrics, we carried out a 2 x 2 repeated-measures ANOVA in which the two levels of emotion category were represented by combining ratings for excerpts intended to convey happiness and calmness, and ratings for excerpts intended to convey sadness and anger. There was still a main effect of lyrics, *F*(1,103)=212.77, *p*<.05, η²G=.15, but the main effect of emotion category was non-significant, *F*(1,103)=.04, *p*=.832. There was, however, a significant interaction between emotion and lyrics, *F* (1,103)=5.81, *p*< .05, η²G < 0.01, illustrated in Figure 2.

[FIGURE 2]

## 

To test Prediction (5), that participants who scored higher on the musical training and emotions subscales of the Gold-MSI would give higher ratings to each emotion category than participants who scored lower on these two subscales, we first calculated Pearson correlation coefficients between participants’ mean ratings of the excerpts intended to convey the four emotions (happy, sad, calm, angry) and their scores on the two subscales. These are presented in Table 1 and illustrated in Figure 3. The variable in the emotion condition refers to the emotion the stimulus was intended to represent and the individual’s rating of the stimulus. For example, happy ratings are each participant’s rating of the item relating to “happy” in relation to the stimulus intended to represent happiness and represent the average rating across an emotional category per participant plotted against their score from the respective subscale of the Gold-MSI. As shown in Table 1 the highest correlations were between the happy and calm ratings and between the musical training and emotions subscales.

[TABLE 1]

[FIGURE 3]

Finally, we used the lmerTest package in R (Kuznetsova et al., 2017) to run a linear mixed-effects model to investigate our prediction that participants who scored higher on the Gold-MSI subscales also gave each emotional category higher ratings. We accomplished this by adding both the emotions and musical training subscales of the Gold-MSI as fixed effects. The results of the mixed-effects model analysis are shown in Table 2. Paralleling the previous analyses that modeled the lyrics condition, this main effect was significant with the no-lyrics condition estimated to be lower (*B*= -0.75) than the lyrics condition. Similarly, our four emotion categories exhibited the same pattern of responses as shown in Figure 2 in addition to the same single significant interaction. Unlike the previously reported ANOVA, the use of our mixed-effects model allowed us to incorporate individual level data for each participants’ Gold-MSI subscale score. This additional information allowed us to investigate whether the musical training or emotion scale made a significant contribution to predicting our response variable. As shown in Table 2, the emotions subscale was a significant predictor (*p* = .03) whereas the musical training subscale was not (*p* = .07). We also performed a similar analysis of the data adding our familiarity rating as a predictor variable to the first mixed-effects analysis. In this analysis, the familiarity rating did not emerge as a significant predictor. This was done to ensure that familiarity with the stimuli did not affect our ratings.

[TABLE 2]

# Discussion

In this study we aimed to replicate and extend Experiment 1 reported by Ali and Peynircioğlu (2006). First, we discuss our replication, then our extension, in which we considered the role of musical training and emotional engagement with music on ratings of excerpts with and without lyrics intended to convey the four emotions, and finally we reflect on the discrepancies between the findings of the original experiment and its replication.

Like Ali and Peynircioğlu (2006) we found main effects of emotion and lyrics on listeners’ ratings of the excerpts, and an interaction between them when ratings of excerpts intended to convey the four categories of emotion (happy and calm, sad and angry) were combined into two categories, positive and negative. Our results differed from those of Ali and Peynircioğlu, however, as our participants awarded lower, not higher, ratings to happy excerpts than to calm, angry and sad excerpts, and the highest ratings to calm excerpts. Our results also differed from those of Ali and Peynircioğlu because excerpts with lyrics received higher, not lower, ratings than excerpts without lyrics, and this affected the interaction between lyrics and emotions (positive vs. negative).

There are several possible explanations for the differences between the two sets of results. First, our experiment was not an exact replication because, although we attempted to obtain the stimuli used in the original study, the authors told us that they were unavailable. The no-lyrics stimuli consisted of pre-existing recordings of jazz and classical melodies to which Ali and Peynircioğlu added lyrics taken from popular songs. It has been proposed by Song et al. (2016) that music in popular and classical genres may elicit different kinds of perceived emotion, as may music composed for different instruments [(Hailstone et al., 2009)](https://paperpile.com/c/eFyHRr/dNDe). Our stimuli consisted of pairs of excerpts from the same songs, half popular tunes and half classical, with and without lyrics. In creating these novel stimuli, we may have introduced confounding variables such as the use of the sung syllable (/nō/) in the no-lyrics excerpts, which could have semantic associations. Second, unlike Ali and Peynircioğlu, who did not specify the basis on which their stimuli were to be rated, we prompted participants to rate the excerpts on the basis of the emotions they perceived rather than the emotions they felt. There is evidence that different mechanisms underlie the experience and perception of emotion, which influence the way they are rated (Gabrielsson, 2001; Juslin, & Västfjäll, 2008). Lack of generalizability has long been noted as a problem in psychology (e.g., Durgin et al., 2012) and is still under discussion (e.g., Yarkoni, 2022); this should be of concern to music psychologists as in all other sub-disciplines of psychology.

A third and perhaps more plausible explanation for the discrepancy between the two sets of findings is that Ali and Peynircioğlu (2006) made a Type I error. With only 32 participants their sample was relatively small, and they tested a large number of variables, which could have produced unstable results. Human error may have played a role, too. We submitted the means reported by Ali and Peynircioğlu (Experiment 1) to a Granularity-Related Inconsistency of Means (GRIM) test (Brown & Heathers, 2017) for the purposes of preparing the figures provided in the present report. This showed that only 7/16 (44%) means could have been obtained given the sample size. Similar inconsistencies were found in the subsequent three experiments reported in the article, with 8/12 (66%), 4/12 (43%), and 7/16 (44%) possible values respectively.

Issues of statistical power and human error aside, the most plausible reason for our failure to replicate the findings of Ali and Peynircioğlu (2006) is that their hypotheses were not sufficiently grounded in theory that would make a specific, directional prediction. This highlights the importance for research to be motivated by theory and the need for verbal reports to reflect the results of statistical testing, if the replication crisis is to be mitigated (Yarkoni, 2022). Future research on listeners’ emotional responses to songs with and without lyrics could be informed by the six mechanisms proposed by Juslin and Västfjäll (2008) or more recent explanations (e.g., Warrenburg, 2020), but most importantly should predict much more specific outcomes.

**Role of individual differences**

Since the publication of Ali and Peynircioğlu’s study in 2006, music psychologists have become increasingly interested in alternative explanations for effects reported in the literature. Musical sophistication (Müllensiefen et al., 2014) could provide one such explanation. Musical training, for example, has been shown to be positively associated with the successful performance of musical tasks, including the decoding of emotional expression (Akkermans et al., 2019; Dahary et al., 2020).

Expanding on our factorial model investigating the effects of emotions, lyrics and gender (Predictions 1-4), we used a linear mixed-effects model to test (5) the predictive effect on ratings of higher scores on the musical training and emotions subscales of the Gold-MSI. When both subscales were included in the model, emotions, but not musical training, made a significant contribution. This may have been a statistical artifact resulting from both subscales reflecting variance that might be caused by a general musical sophistication factor. When the emotions subscale was removed and only the musical training subscale was included in the model, it too made a significant contribution, supporting an effect of musical training for these types of tasks (Song et al., 2016; Vuoskoski & Eerola, 2011). In future research it will be important to disentangle the effects of these variables from each other and from musical sophistication in general.

Based on our findings, we agree that musical training should be measured using tools informed by conceptual distinctions between the ways in which individuals can engage with music (Bigand & Poulin-Charronnat, 2006). Our analysis highlights the importance, in future research, of considering theoretically relevant covariates that are capable of explaining variations in the identification of emotion in music perception.

# Conclusions

In the present study, we replicated Experiment 1 reported by Ali and Peynircioğlu (2006), which examined the effect of lyrics on perceived emotion in music. Some of our results were significant but differed in their specifics from those reported in the original article. We contextualize our results in light of ongoing calls for methodological reform in psychology, noting the need for a wider discussion of generalizability, the need for well-powered studies, and reproducible analyses. In future research on music and emotion we recommend the inclusion of self-report measures such as emotional engagement with music and strongly advocate for the inclusion of directional hypotheses supported by the empirical literature.

# 

# References

Ali, S. O., & Peynircioğlu, Z. F. (2006). Songs and emotions: Are lyrics and melodies equal partners? *Psychology of Music*, 34(4), 511–534. <https://doi.org/10.1177/0305735606067168>

Akkermans, J., Schapiro, R., Müllensiefen, D., Jakubowski, K., Shanahan, D., Baker, D., ... Frieler, K (2019). Decoding emotions in expressive music performances: A multi-lab replication and extension study. *Cognition & Emotion*, 33(6), 1099–1118. <https://doi.org/10.1080/02699931.2018.1541312>

Anglada-Tort, M., & Sanfilippo, K. R. M. (2019). Visualizing music psychology: A bibliometric analysis of Psychology of Music, Music Perception, and Musicae Scientiae from 1973 to 2017. *Music & Science*, 2, 1–18. <https://doi.org/10.1177/2059204318811786>

Baker, D., Ventura, J., Calamia, M., Shanahan, D., & Elliott, E. (2018). Examining musical sophistication: A replication and theoretical commentary on the Goldsmiths Musical Sophistication Index. *Musicae Scientiae*, 24(4), 411–419. <https://doi.org/10.1177/1029864918811879>

Barradas, G. T., & Sakka, L. S. (2021). When words matter: A cross-cultural perspective on lyrics and their relationship to musical emotions. *Psychology of Music*. https://doi.org/ 10.1177/03057356211013390

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, 37, 379–384. <https://doi.org/10.3758/BF03192707>

Bigand, E., & Poulin-Charronnat, B. (2006). Are we “experienced listeners”? A review of the musical capacities that do not depend on formal musical training. *Cognition*, 100(1), 100–130. <https://doi.org/10.1016/j.cognition.2005.11.007>

Brattico, E., Alluri, V., Bogert, B., Jacobsen, T., Vartiainen, N., Nieminen, S. K., & Tervaniemi, M. (2011). A functional MRI study of happy and sad emotions in music with and without lyrics. *Frontiers in Psychology*, 2, 308. <https://doi.org/10.3389/fpsyg.2011.00308>

Bullack, A., Büdenbender, N., Roden, I., & Kreutz, G. (2018). Psychophysiological responses to “happy” and “sad” music: A replication study. *Music Perception*, 35 (4), 502–517. <https://doi.org/10.1525/mp.2018.35.4.502>

Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, 8(4), 363–369. <https://doi.org/10.1177/1948550616673876>

Castro, S. L., & Lima, C. F. (2014). Age and music expertise influence emotion recognition in music. *Music Perception*, 32 (2), 125–142. <https://doi.org/10.1525/mp.2014.32.2.125>

Chamorro-Premuzic, T., Gomà-i-Freixanet, M., Furnham, A., & Muro, A. (2009). Personality, self-estimated intelligence, and uses of music: A Spanish replication and extension using structural equation modeling. *Psychology of Aesthetics, Creativity, and the Arts*, 3(3), 149–155. <https://doi.org/10.1037/a0015342>

Dahary, H., Palma Fernandes, T., & Quintin, E. M. (2020). The relationship between musical training and musical empathizing and systemizing traits. *Musicae Scientiae, 24(1)*, 113–129. <https://doi.org/10.1177/1029864918779636>

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47(1), 1–12. <https://doi.org/10.3758/s13428-014-0458-y>

Fiveash, A., Luck., L. (2016). Effects of musical valence on the cognitive processing of lyrics. *Psychology of Music*, 44 (6), 1346–1360. <https://doi.org/10.1177/0305735615628057>

Frieler, K., Müllensiefen, D., Fischinger, T., Schlemmer, K., Jakubowski, K., & Lothwesen, K. (2013). Replication in music psychology. *Musicae Scientiae*, 17(3), 265–276. <https://doi.org/10.1177/1029864913495404>

Gabrielsson, A. (2001). Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, 5, 123–147. <https://doi.org/10.1177/10298649020050S105>

Hailstone, J. C., Omar, R., Henley, S. M. D., Frost, C., Kenward, M. G., & Warren, J. D. (2009). It’s not what you play, it’s how you play it: Timbre affects perception of emotion in music. *Quarterly Journal of Experimental Psychology*, 62(11), 2141–2155. <https://doi.org/10.1080/17470210902765957>

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., & Scheibe, S. (2014). Multcomp: simultaneous inference in general parametric models. R Package Version, 1–3. <http://multcomp.r-forge.r-project.org/>

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science*, 359(6377), 725–726. <https://doi.org/10.1126/science.359.6377.725>

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>

Ioannidis, J. P. A. (2016). Why Most Clinical Research Is Not Useful. *PLoS Medicin*e, 13(6), e1002049. <https://doi.org/10.1371/journal.pmed.1002049>

Juslin, P., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, 31(5), 559–575. <https://doi.org/10.1017/S0140525X08005293>

Karreman, A., Laceulle, O. M., Hanswer, W, E., & Vingerhoets, A. (2017). Effects of emotion regulation strategies on music-elicited emotions: An experimental study explaining individual differences. *Personality and Individual Differences*, 114 (1), 36–41. <https://doi.org/10.1016/j.paid.2017.03.059>

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82(13)*, 1–26. <https://doi.org/10.18637/jss.v082.i13>

Lawrence, M. (2016). ez: Easy analysis and visualization of factorial experiments. R package version 4.4-0. <https://CRAN.R-project.org/package=ez>

Lin, H.-R., Kopiez, R., Müllensiefen, D., & Wolf, A. (2019). The Chinese version of the Gold-MSI: Adaptation and validation of an inventory for the measurement of musical sophistication in a Taiwanese sample. *Musicae Scientiae*. <https://doi.org/10.1177/1029864919871987>

[REDACTED]. (2019, August). *Lyrics and Emotion in Songs: A Conceptual Replication Study of Ali and Peynircioglu, 2006.* [REDACTED]

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PloS One*, 9(2), e89642. <https://doi.org/10.1371/journal.pone.0089642>

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological Methods*, 8(4):434–47. <https://doi:10.1037/1082-989X.8.4.434>

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251). <https://doi.org/10.1126/science.aac4716>

Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060), 1226–1227. <https://doi.org/10.1126/science.1213847>

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161. <https://doi.org/10.1037/h0077714>

Song, Y., Dixon, S., Pearce, M. T., & Halpern, A. R. (2016). Perceived and induced emotion responses to popular music: categorical and dimensional models. *Music Perception*, 33 (4), 472–492. <https://doi.org/0.1525/mp.2016.33.4.472>

Susino, M., & Schubert, E. (2019). Cultural stereotyping of emotional responses to music genre. *Psychology of Music*, 47(3), 342–357. <https://doi.org/10.1177/0305735618755886>

Susino, M., & Schubert, E. (2020). Musical emotions in the absence of music: A cross-cultural investigation of emotion communication in music by extra-musical cues. *PLoS ONE*, 15(11), Article e0241196. <https://doi.org/10.1371/journal.pone.0241196>

Taruffi, L., Allen, R., Downing, J., & Heaton, P. (2017). Individual differences in music-perceived emotions: The influence of externally oriented thinking. *Music Perception*, 34 (3), 253–266. <https://doi.org/10.1525/mp.2017.34.3.253>

Tsai, C.-G., Chen, R.-S., & Tsai, T.-S. (2014). The arousing and cathartic effects of popular heartbreak songs as revealed in the physiological responses of listeners. *Musicae Scientiae*, 18(4), 410–422. <https://doi.org/10.1177/1029864914542671>

​​Vuoskoski, J. K., & Eerola, T. (2011). Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences. *Musicae Scientiae*, 15(2), 159–173. <https://doi.org/10.1177/1029864911403367>

Warrenburg, L. A. (2020). Comparing musical and psychological emotion theories. *Psychomusicology: Music, Mind, and Brain*, 30(1), 1–19. <https://doi.org/10.1037/pmu0000247>

Weth, Karim., Raab, H. Marius., Carbon, Claus-Christian. (2015). Investigating emotional responses to self-selected sad music via self-report and automated facial analysis. *Musicae Scientiae*, 19(4), 412-–432. <https://doi.org/10.1177/1029864915606796>

Wolf, A., Kopiez, R., Platz, F., Lin, H.-R., & Mütze, H. (2018). Tendency towards the average? The aesthetic evaluation of a quantitatively average music performance: A successful replication of Repp’s (1997) study. *Music Perception*, 36(1), 98–108. <https://doi.org/10.1525/mp.2018.36.1.98>

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. <https://doi.org/10.1017/S0140525X20001685>

1. Calculations based on citations reported in Anglada-Tort and Sanfillipo (2019). [↑](#footnote-ref-1)
2. Here we adopt Nosek and Errington’s [(2020)](https://paperpile.com/c/eFyHRr/5coa/?noauthor=1) definition of replication as “a study for which any outcome would be considered diagnostic evidence about a claim from prior research” (p. 2), which they discuss at length. [↑](#footnote-ref-2)
3. Although at the time of our pre-registration, we did not include an intent to study the effects of gender, we include those analyses here in order to engage with the original analyses. [↑](#footnote-ref-3)
4. URL: [REDACTED FOR PEER REVIEW] [↑](#footnote-ref-4)
5. We represent effect size using generalized eta squared (η²G ) as it is useful for comparing between- and repeated- measures designs, particularly for the purposes of meta-analysis (Bakeman, 2005; Olejnik & Algina, 2003). [↑](#footnote-ref-5)