Running Head: EFFECT OF LYRICS ON EMOTION

Generalizing the Effect of Lyrics on Emotion Ratings

Yiqing Ma[12], David John Baker[23], Katherine M. Vukovics[24], Connor J. Davis[25], Emily M. Elliott[2]

[1] University of Michigan

[2] Louisiana State University

[3] Goldsmiths, University of London

[4] Anderson High School

[5] John Brown University

Author Note

Yiqing Ma, Department of Music Theory, School of Music, Theater and Dance, University of Michigan, Ann Arbor, MI; David John Baker, Department of Computing, Goldsmiths, University of London, London, United Kingdom; Connor J. Davis, John Brown University, Siloam Springs, AR. Katie M. Vukovics, Associate Choir Director, Anderson High School, Austin, TX. Emily Elliott, Department of Psychology, Louisiana State University, Baton Rouge, Louisiana.

Data from this experiment were presented at the Psychonomic Society's 61st Virtual Meeting, November, 2020.

Corresponding Author: Emily Elliott, Department of Psychology, 236 Audubon Hall, Louisiana State University, Baton Rouge, LA, USA 70803, eelliott@lsu.edu

Running Head: EFFECT OF LYRICS ON EMOTION

Abstract

What factors affect listeners' emotional perception of music? Ali and Peynircioğlu (2006) conducted a series of experiments on the listener's emotional response to melodies and lyrics in songs. Here we present a pre-registered replication and extension of this line of research with new adapted stimuli and several musical covariates using the Goldsmiths Musical Sophistication Index (Gold-MSI). Using a within-subjects design, participants (n = 104) were asked to rate the perceived emotions to unfamiliar happy, sad, calm, and angry songs with and without lyrics to model the extent to which each factor contributed to listener ratings. We failed to replicate the results of the original paper, but did find significant results in the opposite direction of several variables. Our extension of the analysis found evidence supporting the idea that emotional perception can and should be divorced from aspects associated with musical training. The mixed-effects model showed a significant effect of the emotional engagement subscale of the Gold-MSI above and beyond our replication analysis. While the results we detail in this research report conflict with the original findings, this conceptual replication and extension serve to highlight the importance of replicating findings reported within the music psychology literature.

Keywords: song, emotion, melody, lyrics, circumplex model

Running Head: EFFECT OF LYRICS ON EMOTION

Generalizing the Effect of Lyrics on Emotion Ratings?

According to Google Scholar, the paper "Songs and emotions: Are lyrics and melodies equal partners?" (Ali & Peynircioğlu, 2006) has been cited over 240 times since its initial publication. Supplementing this metric with a recent bibliometric analysis examining the music psychology literature, this paper places near the top ten percent of papers cited within music psychology (Anglanda-Tort & Sanfillipo, 2019).[1] In the paper, Ali and Peynircioğlu investigated the degree to which having lyrics present or absent in a song affected the ratings of various emotions given by a listener. Understanding the extent that lyrics affect various emotional responses establishes what subsequent researchers might expect in similar designs (Fiveas & Luck, 2016; Brattico et.al, 2011)

In this paper we report findings from a replication experiment with a design similar to Ali and Peynircioğlu (2006), but with increased sample size and pre-registered analyses in order to investigate the extent that the relationship between lyrics and emotions generalizes to stimuli beyond those initially tested in the first paper. We also investigate the degree that this effect generalizes by collecting covariates that might explain variation in the data given literature published since the initial publication concerning individual differences (e.g., Akkermans et al., 2019; Dahary, Palma Fernandes, & Quintin, 2020; Weth, Raab, & Carbon, 2015).

Replication in Music Psychology

As a subset of psychology, music psychology has not been immune to what is referred to as the "replication crisis" (Ioannidis, 2005; Ioannidis, 2016; Hutson, 2018; Peng, 2011; Open Science Collaboration, 2015). Frieler and colleagues argue that findings in music psychology

[1] Calculations based on citations reported in Anglada-Tort and Sanfillipo (2019)

might be even more unstable than other subspecialties of psychology because music psychology is “low-gain/low-cost science” (Frieler et al., 2013; page 271). They argue that due to the general lack of resources that result in smaller sample sizes in addition to the relative harmlessness of theories related to music psychology, research in music psychology might tend to value novelty, creativity, or originality over validity and reliability. Music psychology has since begun publishing more replications of studies ranging from psychometrics (Baker et al., 2018; Chamorro-Premuzic et al., 2009; Lin et al., 2019) , physiological responses to music (Bullack et al., 2018), and the evaluation of musical performances (Wolf et al., 2018).

Replicating Findings

Ali and Peynircioğlu (2006) has served as an important reference for subsequent papers on emotion perception in music (Tsai, Chen, Tsai, 2014; Brattico et. al, 2011). The paper reports four different experiments, each asserting several significant findings. In this research report, we focus on the first of the four experiments.

The first experiment (N = 32) examined the emotional responses to music with and without lyrics, crossing the lyric condition with music from four emotional categories: 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 for their methods and design can be found in the original manuscript (Ali & Peynircioğlu, 2006).

[FIGURE 1]

The authors reported a main effect of the emotional category of the musical excerpt, with only the category of happy being significantly different from the other three emotions using post-hoc comparisons, as depicted in Figure 1A. They also reported a main effect of lyrics, with melodies without lyrics scoring higher than those with lyrics (Figure 1B). While the authors did not report a significant effect of gender ($F(1, 30) = 0.30$, $MSE = 3.81$, $p > .10$, not shown in Figure 1), they reported a significant interaction of lyrics and gender, with women rating the emotions associated with the melody-only condition lower than men ($F(1, 30) = 4.57$, $MSE = 1.13$, $p < .05$, not shown in Figure 1). The authors further conducted a separate analysis that grouped positive (happy and calm) and negative (sad and angry) emotions together and tested these combined and reported that positive music was rated higher than negative music in all categories (Figure 1C).

In the current study, we took the following approach. First, we replicated their statistical analysis as reported in the original paper. We note that this is not a direct replication in that we were unable to obtain the original materials and thus created our own stimuli. Further, we asked participants to make their ratings of the music regarding perceived emotion as this detail was not explicitly addressed in the original publication.

Second, given literature suggesting that individual differences such as musical training (Akkermans et al., 2019; Dahary, Palma Fernandes & Quintin, 2020) might explain variation in responses, we extended the original replication analysis by also collecting individual data using the Goldsmiths Musical Sophistication Index (Müllensiefen, Gingras, Musil, & Stewart, 2014) to provide a robust, continuous measure of musical sophistication. We did this to provide novel insights into the literature on music and emotion perception.

To organize our analyses between our replication and extension, we provided the following two hypotheses: (1) in our replication analysis, we predicted we would directly replicate the analysis presented in the original paper, as detailed above. This hypothesis included a main effect of lyrics and emotion in the direction initially reported as well as the significant interaction, in addition to the grouped analysis. (2) In our extension analysis, we predicted that participants who scored higher on the Gold-MSI musical training and emotional response subscale would lead to higher ratings than those with lower scores on each of these subscales. We modeled our emotional ratings presuming the inclusion of these fixed effects would result in adjusted means for our dependent variable.

All hypotheses, design, and analyses were pre-registered through Aspredicted.org and all materials are hosted on the Open Science Framework (OSF)[2]. A pilot version of the study was run in the Summer of 2019 ([REDACTED], 2019) which provided support for the use of the current materials.

Method

Participants

One hundred and eight participants were recruited for this experiment at a University in the Southeastern United States. All participants were awarded extra credit or course credit in undergraduate psychology or music theory courses. Participants (n=74) were recruited from the Department of Psychology and the School of Music (n=34). All included participants are native speakers of American English and have reported having normal hearing and vision. These

[2] URL: https://osf.io/dzqx3/?view_only=ac6bb994c822421caa45d5c3e2ad3c77

exclusion criteria removed four participants resulting in a final sample of n = 104. The final sample comprised 62 women and 42 men (M_{age} = 19.2, SD_{age} = 1.37, R_{age} = 18-26).

Materials

As mentioned above, we attempted to obtain the original musical selections from the authors, but the stimuli were unavailable per correspondence with the original authors. We created a new set of stimuli based on the original materials. A total of sixteen unfamiliar songs were chosen from different musical genres to elicit four different types of emotional response, corresponding to the four quadrants of Russell's (1980) circumplex theory of emotion (happy, sad, calm, angry). These unfamiliar stimuli were chosen by songs with relatively low playtimes in Spotify (see Supplemental Material). Our pilot study provided evidence that the songs we chose could be expected to be unfamiliar to our participant pool and also conveyed the intended emotional category we intended. We recorded 35-second clips containing the first verse or refrain of songs.

Unlike the original paper, each song originally had lyrics that were subsequently removed and recorded using a MIDI keyboard accompanying a singer who either sang the lyrics or sang on a neutral syllable "no (/nō/)" to create the no-lyrics complement of the song. This methodological change from the original design was made to ensure consistency of the melody between the melody only and melody and lyrics conditions.

Design

The thirty-two songs were then divided into two blocks, with equal numbers of lyrics and no-lyrics songs in each block. 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.

After listening to each song 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 make five ratings on a nine-point Likert scale. The first four ratings corresponded to their happy, calm, sad, and angry ratings. The fifth rating asked participants how familiar they were with the music they just heard.

After completing the listening task, participants were asked to complete the Goldsmith Musical Sophistication Index (Müllensiefen et. al, 2014). Participants provided information related to musical training experiences, musical listening habits, and demographic information.

Both sessions took place in a sound-attenuating environment. Stimuli were presented to participants in jsPsych (de Leeuw, 2015). Stimuli were presented via Sennheiser HD 229 over-the-ear headphones, connected with a Windows 10 PC (volume was judged to be subjectively comfortable by the research team).

Results

The analysis plan here follows our pre-registration #19280 filed in the Winter of 2019 on Aspredicted.org that can be found on this project’s OSF repository.

Hypothesis 1: Replication

To investigate our first research question, 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) corrected with Greenhouse-Geisser to examine the influence of three independent variables (lyrics,

emotions, gender) on ratings of emotion. Our lyrics condition contained two levels (lyrics, no-lyrics), our emotions condition contained four levels (happy, sad, calm, angry), and our between-subjects gender variable contained two levels (female and male). The ANOVA was computed with Type III sum-of-squares.

We report only main effects of gender and lyrics, with no significant main effect of gender or any interactions. The main effect of emotions yielded an $F(3, 264.30) = 28.24$, $p < .001$, $\eta^2_G = .11$ indicating a significant difference between happy ($M = 5.10$, $SD = 2.71$), sad ($M = 6.06$, $SD = 2.52$), calm ($M = 6.63$, $SD = 1.98$), and angry ($M = 5.56$, $SD = 2.98$) emotions.[3] The main effect of lyrics yielded an $F(1, 103) = 203.46$, $p < .001$, $\eta^2_G = .10$ indicating a significant difference between ratings of songs with lyrics ($M = 6.34$, $SD = 2.52$) and without lyrics ($M = 5.34$, $SD = 2.64$). No other effects were significant (ps > .05).

We followed up on our main effects of both Emotions and Lyrics using a Tukey HSD test implemented in the multcomp package in R (Hothorn et al., 2014). We report significant differences ($p < .05$) in the Emotion condition between all pairs of emotions except sad and calm. We also report a significant pairwise difference between the lyrics condition ($p < .05$) as shown in Figure 2.

To replicate the grouped analysis as reported in the original paper, we grouped together the happy and calm conditions with the sad and angry conditions to create a variable with two levels (negative and positive) in addition to the lyrics factor resulting in a 2 x 2 repeated measures ANOVA. There was no main effect $F(1,103) = .04$, $p = .832$ between negative ($M = 5.83$, $SD = 2.77$) and positive emotions ($M = 5.86$, $SD = 2.49$). The main effect of Lyrics

[3] We present generalized Eta Squared (η^2_G)as it is easier to compare between and repeated measures designs, thus more conducive to use in meta analysis (Bakeman, 2005; Olejnik & Algina, 2003)

remained significant $F(1,103) = 212.77$, $p < .05$, $\eta^2_G = .15$ and we report a significant interaction $F(1,103) = 5.81$, $p < .05$, $\eta^2_G < 0.01$.

[FIGURE 2]

Hypothesis 2: Extension

To investigate our second research question of musical training, we first conducted a correlational analysis between each emotional condition and both the musical and emotional subscales. Data for each correlational analysis are presented in Table 1 and in Figure 3 and represent the average rating across an emotional category per participant plotted against their score from the respective subset of the Gold-MSI.

[TABLE 1]

[FIGURE 3]

To model our extension analysis, we ran a linear mixed-effects model predicting our emotions rating, having added both the emotions and musical training factors as fixed effects using the lmerTest package in R (Kuznetsova, Brockhoff, & Christensen, 2017). Results of the mixed-effects model are listed in Table 2. Results from the mixed-effects model were overall very similar to the replication analysis with both main effects of lyrics and emotion. In addition to these variables being significant, the Emotional Engagement subscale of the Gold-MSI

yielded a significant result ($p < .05$) as did an interaction between the no-lyrics and sad music conditions.

[TABLE 2]

Discussion

The goal of this paper was to replicate and extend the findings of the first experiment reported by Ali and Peynircioğlu (2006). We first discuss our attempts at replication, then move to our extension, and finally end with a reflection on the discrepancies between our findings and the original paper.

Replication

Our first goal was to replicate the analysis reported in the first experiment of Ali and Peynircioğlu (2006). Specifically, we were interested in comparing the pattern of responses in the main effects, interaction, and grouped analysis. As in the original experiment, we report a main effect of emotion, a main effect of lyrics condition as well as a significant interaction, but in the opposite direction of the original paper.

Regarding the main effect of emotion, the original study reported that only happy songs were significantly different from all three other emotions when subjected to post-hoc comparisons. While we did find that songs from the happy category were significantly different from the other categories, we additionally found significant differences between all pairs of emotional pairings with only the two negative emotions of sad and angry not differing from one another as shown in Figure 2. More importantly, our calm songs were rated as the highest of the

four, with happy songs receiving the lowest of all the ratings. This outcome is reversed from the original study and suggested that under the current context, the initial findings were not robust and did not generalize. Similarly, we found a reversal of direction in our lyrics and no-lyrics condition and were not able to replicate the grouped analysis.

In considering the difference in findings, we posit several plausible explanations. The first and major consideration in interpreting our reported findings is that our experiment was not an exact replication of the original study. 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 (Gabrielsson, 2001; Juslin, & Västfjäll, 2008). Further, these generalizability problems have been the subject in ongoing discussions of psychological reform (Yarkoni, 2020). We believe these discussions must also be considered within music psychology.

A second and more practical explanation for the findings might be that the initial findings of Experiment 1 of Ali and Peynircioğlu (2006) were a Type I error. The relatively small sample size of the initial paper (N = 32) may have led to unstable results given the number of variables tested. A third, and more technical explanation for the lack of replication might be human error in documenting these analyses. In preparing the data used in the figures in this manuscript, our team submitted the reported means of the tables in each of the reported experiments to a Granularity-Related Inconsistency of Means (GRIM) test (Brown and Heathers 2017).

When submitted to the GRIM test, the reported means for Experiment I indicate that only 44% (7/16) of the reported means are numerically possible given the sample size of 32

participants. Similar inconsistencies were found in the other three experiments reported in the paper which respectively have 66% (8/12), 43% (4/12), and 44% (7/16) possible values.

Issues of statistical power and human error aside, the most probable reason for the lack of ability to replicate these findings would be the lack of theoretical support for the hypotheses predicting a specific outcome leading the initial results to be over-interpreted. We believe this possibility highlights the need for music psychologists to continue to improve the theoretical motivations they choose to investigate with statistical tests (Yarkoni, 2020).

Concerning generalizability, our choice to use modified stimuli may have also affected the results of our replication. The original study opted to use various jazz and classical instrumental tracks for their no-lyrics condition. Our replication study instead used the same songs with and without lyrics to better infer what might be responsible for the pattern of our results. Whereas their choice of stimuli could be argued to be more ecologically valid, research suggested it may be confounded since classical music and popular music have been shown to elicit different emotions (Song, Dixon, Pearce, & Halpern, 2016), a finding also reported with different instrumentations (Hailstone et.al, 2009).

Extension

Since the initial publication of Ali and Peynircioğlu (2006) the music psychology community has adopted more of an interest in individual differences measures to possibly explain effects reported in the literature. One presumably important covariate that music psychologists often consider are aspects related to musical sophistication as summarized in Müllensiefen et al. (2014). As discussed above, measures of musical training have been shown to

be positively associated with various aspects of musical performance tasks, including work on emotional recognition (Akkermans et. al, 2019; Dahary, Palma Fernandes, & Quintin, 2020).

Expanding on our factorial model investigating the musical features and gender, our second analysis incorporated both the musical and emotional subscales of the Gold-MSI. In this model, when both the musical and emotional subscales were added to the model, the emotional fit subscale contributed significantly to the model. While the musical training factor did not contribute significantly, this may be a statistical artifact attributed to the collinearity between the musical training variable and the emotional engagement variable. If this was the case, there is literature to support the effect of musical training for these types of tasks (Vuoskoski & Eerola, 2011; Song, Dixon, Pearce, & Halpern, 2016). Disentangling the effect of emotional engagement from musical training is a future problem in need of addressing for similar research.

We believe our findings corroborated claims encouraging measuring aspects of music perception differently from aspects of music production (Bigand & Poulin-Charronnat 2006). We consider this a point of interest in our analysis as it highlights the importance of future work in music psychology to consider theoretically relevant covariates that are capable of explaining variation in response data in musical identification tasks.

Conclusions

In this paper, we attempted to replicate the first experiment from a paper by Ali and Peynircioğlu (2006) examining the effect of emotions and lyrics. We reported significant findings but reported effects from several variables in the opposite direction of the original paper. Although we are not able to report a successful generalization of the findings, we contextualize these findings in light of ongoing methodological reform in psychology noting the need for a

larger discussion on generalizability, the need for well-powered studies, and reproducible analyses. In extending work in music and emotions we find support for the inclusion of task-relevant individual differences such as emotional engagement with music.

References

Ali, S. O., & Peynircioğlu, Z. F. (2006). Songs and emotions: Are lyrics and melodies equal partners? *Psychology of Music*, 34(4), 511–534. doi: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. doi: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.* doi: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. doi: 10.1177/1029864918811879

Bakeman, R. (2005). Recommended Effect Size Statistics for Repeated Measures Designs. Behavior Research Methods. 37, 379–384. 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. doi: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. doi: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. doi: 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

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. doi: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. doi: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. doi: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

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. doi: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. doi: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.

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. doi: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. doi:10.1017/S0140525X08005293

Kuznetsova A, Brockhoff PB, Christensen RHB (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82(13)*, 1–26. doi: 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*. doi: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, Stephen, and James Algina (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. doi:10.1126/science.1213847

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing [software program], Vienna Austria. http://www.r-project.org

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161. doi: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. doi: 10.1525/mp.2016.33.4.472

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. doi: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

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. doi: 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. doi:10.1525/mp.2018.36.1.98

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. doi:10.1017/S0140525X20001685

Tables

Variable	1	2	3	4	5	6
1 Musical	-					
2 Emotion	.47**	-				
3 Happy	.10	.14	-			
4 Sad	.05	.16	.32**	-		
5 Calm	.30**	.33**	.48**	.33**	-	
6 Angry	.34**	.22**	.04	.36**	.16	-

Table 1: Correlations between Gold-MSI Sub-Scales and average ratings across participants. *$p < .05$. ** $p < .01$.

Predictors	Estimates	CI	*p*
(Intercept)	3.78	[2.62 - 4.95]	<.001
Condition:Melody	-0.75	[1.06 -- -0.44]	<.001
Condition: Calm	1.53	[1.25 -- 1.90]	<.001
Condition: Angry	0.64	[0.34 -- 0.95]	<.001
Condition: Sad	1.24	[0.93 -- 1.55]	<.001
Musical Training	0.01	[0.00 -- -0.03]	.065
Emotional Engagement	0.04	[0.00 -- -0.08]	.028
Melody x Calm	-0.13	[-0.56 -- 0.31]	.560
Melody x Angry	-0.36	[-0.80 -- 0.07]	.100
Melody x Sad	-0.50	[-0.93 -- -0.07]	.024
Random Effects			
Variance	5.80		
ICC	0.08		
N Subjects	104		
Observations	3776		
Marginal R2	.097		
Conditional R2	.172		

Table 2: Mixed effects model incorporating replication analysis variables in addition to musical and emotional engagement subscale from Gold-MSI

Figures

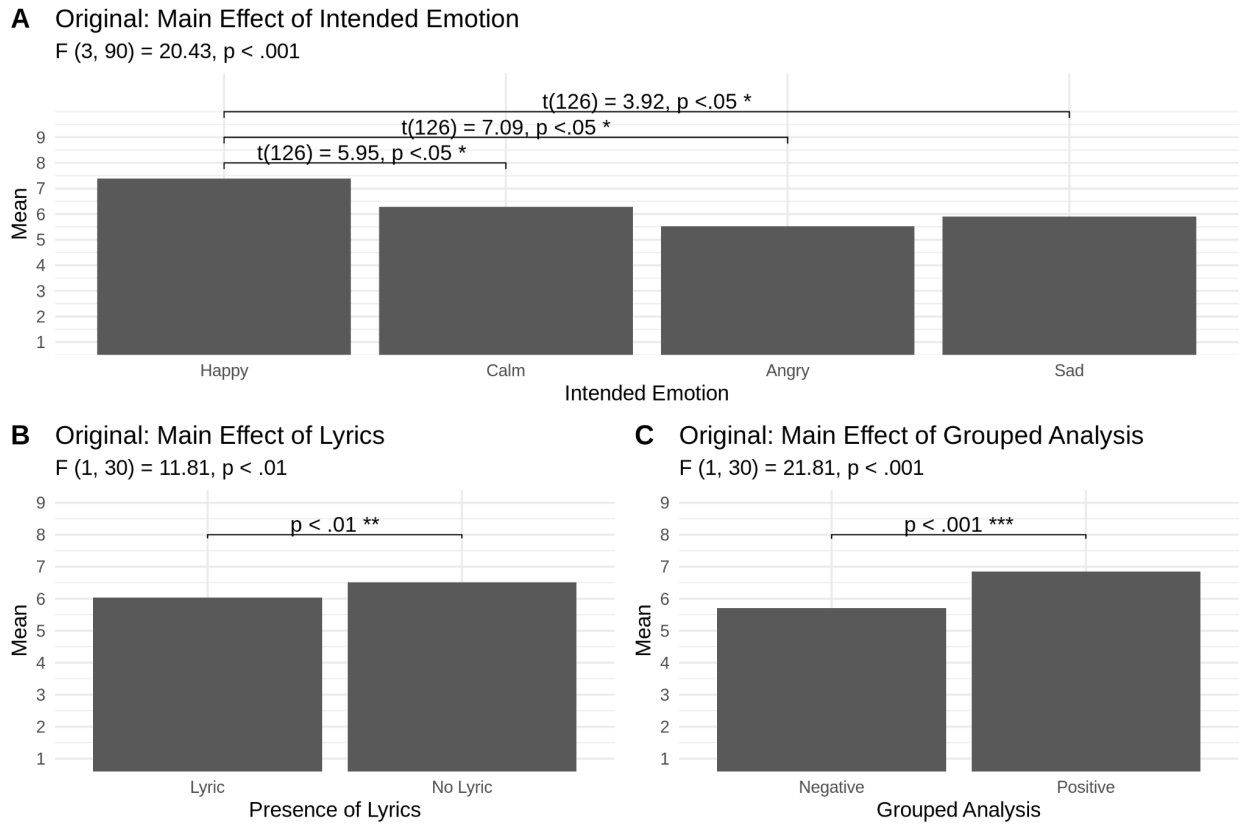

Figure 1: Recreation of Experiment 1 significant results from Ali and Peynircioğlu (2006). Error bars are not included as no error was reported in the original paper's Table 1.

Running Head: EFFECT OF LYRICS ON EMOTION

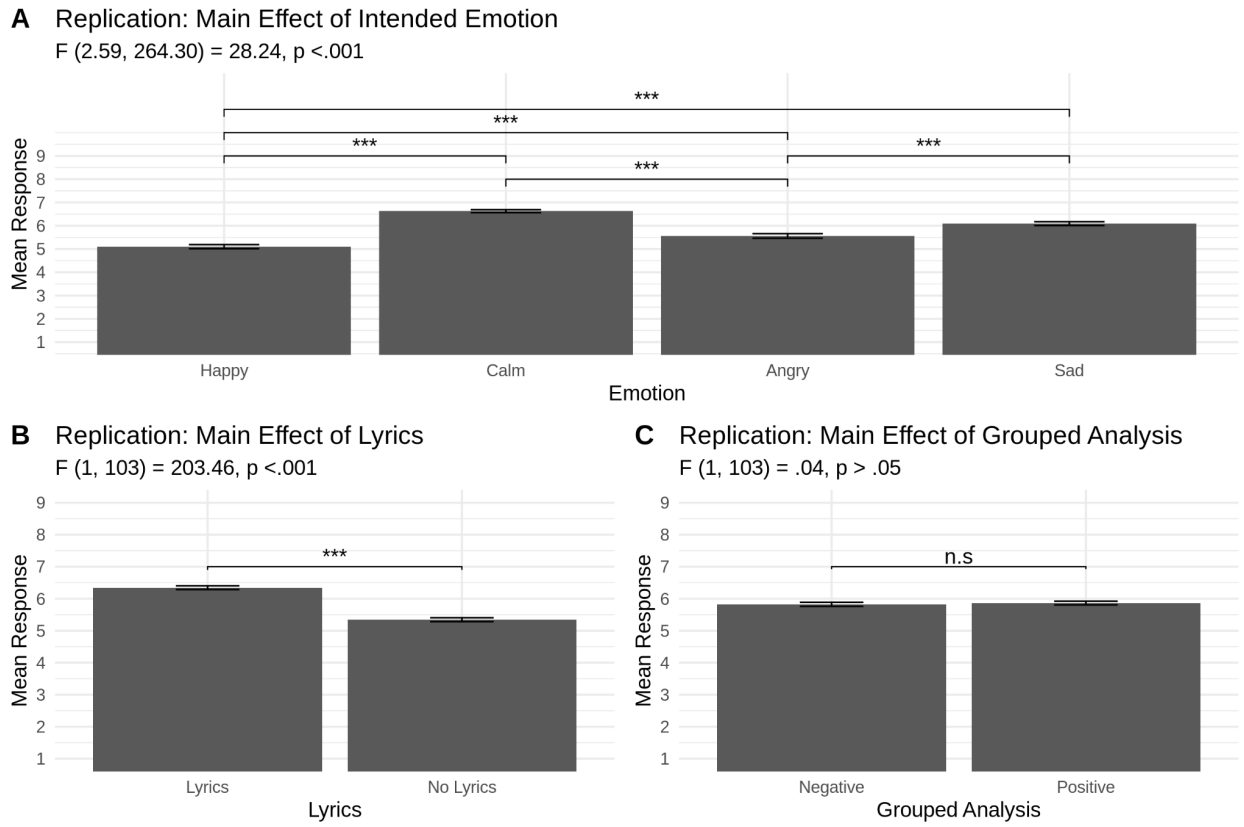

Figure 2: Results of replication of Experiment 1 with replication sample, Results shown here are were the significant main effects reported in the original experiment

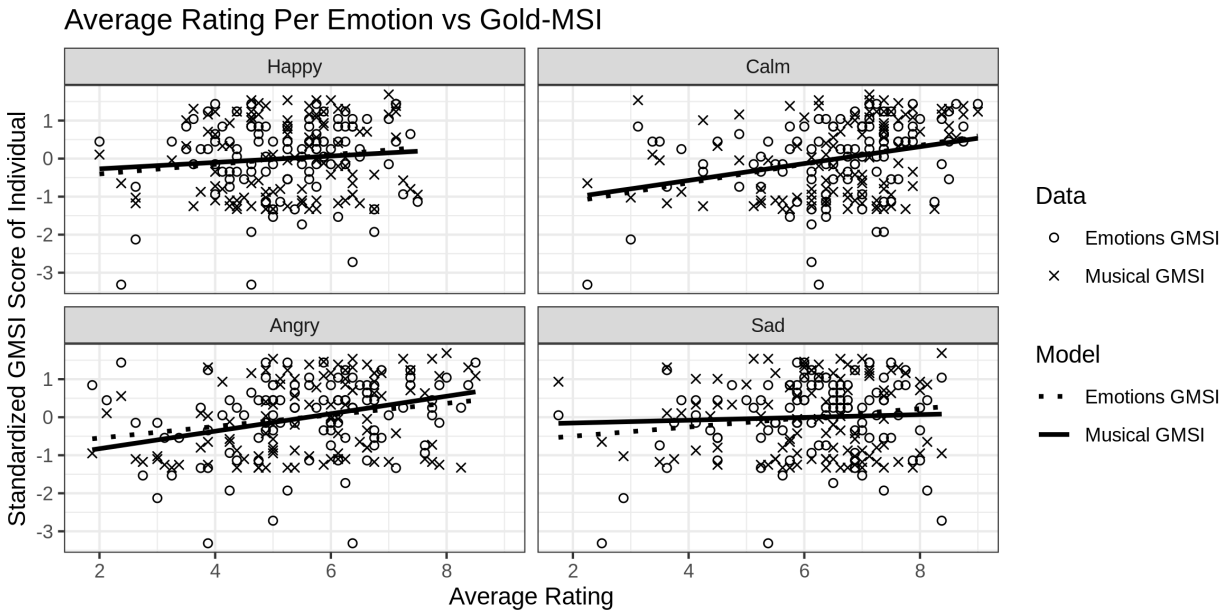

Figure 3: Visualizations of Correlations Reported in Table 1