**Does Exposure to High-Valence Music Cause an Increase to Perceived Mood?**

Tom Ashkenazi, Tres Pimentel, Robert Turnage

**Introduction**

Listening to music has been shown to have many positive effects: increasing self-reported levels of happiness[[1]](#footnote-0), lowering stress levels, lowering anxiety, as well as many other positive effects[[2]](#footnote-1). Anecdotally, many people enjoy listening to music, and while they do so for many different reasons, it seems that people listen to music primarily because they enjoy doing so, that is, they get some positive mood benefit out of it.

Music preference varies widely from person to person. How people settle on their favorite songs or favorite genre of music is a heavily studied phenomenon. What is clear is that listening to your favorite music can provide emotional benefits.[[3]](#footnote-2) [[4]](#footnote-3)

Many studies revolving around people’s reactions to music involve either self-selection of music by participants, or measurement of some outcome variable(s) other than mood (although many of these outcome variables could be used as proxies for mood, or could themselves be described as more descriptive/informational than mood)

We, in this study, are more concerned with people’s reactions to music they have not selected, that is, music that is rather being chosen for them. To explicitly state the question we are interested in answering: we attempted to find whether exposure to happy-sounding music increases self-perceived mood level. As stated previously, happy music can lower stress and anxiety. Music can also change how people perceive the world around them - happy music seems to cause those exposed to it to experience a “happier” perception of the world[[5]](#footnote-4). We intend to show a similar, broader case of what is displayed in these studies - by measuring mood.

We believe that there is enough evidence to reasonably believe that our hypothesis could be true. Based on our methodology, power analysis, and these prior studies, we felt we had a good shot at examining this problem from a slightly different lens than we have seen before - using a very short exposure time as well as utilizing Spotify audio metrics to look at happy music with a new perspective.

**Experimental Details**

*Experiment Design*

Broadly, the design of our experiment was as follows. Subjects were recruited through either convenience sampling or through posts on social media (Reddit, Instagram, Linkedin). Upon clicking the link, subjects were brought to the survey and were given information that we (identified as 3 graduate students from the University of California, Berkeley) were conducting a music psychology-related study. All participants had to check a box indicating that they gave us permission to use their data.

Subjects then were brought to the first survey (pre-listening survey) where they were asked several questions, including the first question to measure treatment effect. After completing this, they were prompted to listen to 3 30-second song snippets (subjects had to navigate to each page, and could leave the page at any time) and then finally, a second survey (post-listening survey) that asked a few more questions, including the second question to measure treatment effect. Once they finished the second survey, they were informed that they could now close the page and leave at any time.

The two questions to measure treatment effect were: “How do you feel right now?”. Participants were prompted to give answers on a 5 point Likert scale.

*Comparison of Potential Outcomes*

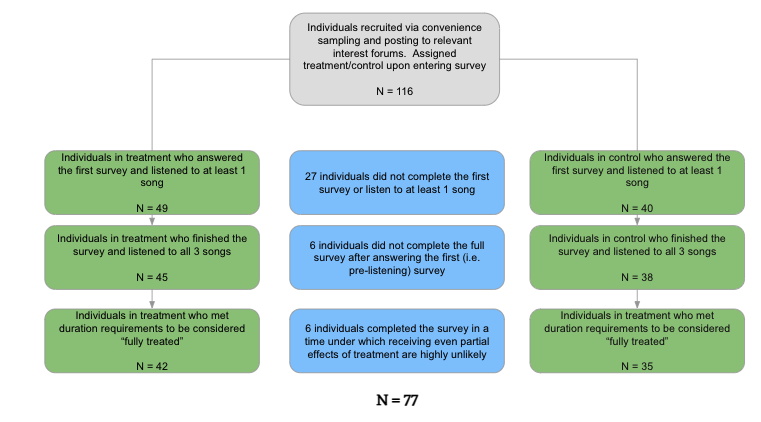
In both our treatment and control groups, we conducted two mood measurements. Both groups self-assessed mood on a 5 point Likert scale upon entering the survey, and then listened to either the treatment or placebo song snippets, and then once again self-assessed mood on the same 5 point Likert scale (i.e. treatment shape is ROXO and control shape is RO-O). The outcome variable we are directly measuring is the difference in these mood self-assessments.

**Formula 1**

moodΔ = moodpost – moodpre

*Sample Size - CONSORT*

**Figure 1: CONSORT Table**



We experienced substantial non-compliance in our experiment. We considered different methods of measuring treatment effects and conducting analysis given this non-compliance. We discuss our methodology and considerations further in the Non-Compliance section. We excluded from the study entirely the 27 individuals who did not complete the survey or listen to at least 1 song. The vast majority of these individuals had the survey page open for under 15 seconds and did not answer a single question - in fact, it is possible they did not even advance past the landing page itself. Because of this, we discarded these observations entirely. Therefore, as we will further detail in the Results section, we looked at both the ITT (using the middle green boxes with n = 83) and the CACE (using the bottom green boxes with n = 77).

*Randomization Process*

Qualtrics has a built-in process for management of the randomization within a survey. To start, we randomized the assignment of treatment or placebo. Then, once assigned to treatment or control, the participant received a random selection of 3 songs (without replacement) from either the treatment or control playlist (each are 10 songs long). Once a participant enters the survey and agrees to participate past the pre-survey questions, the Qualtrics randomization block is used to make the selections and the system records the outcomes. Data on the specific layout of the survey in Qualtrics is available upon request.

*Treatment*

Both groups listened to three 30 second song snippets. The treatment group listened to snippets randomly chosen from a list of 10 “high-valence” songs we specifically picked, while the control group listened to snippets randomly chosen from a list of 10 “low-valence” songs we specifically picked.

Valence, in this context, is a metric measured by Spotify and assigned to all songs on the platform. They define valence as:

“A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).”

One of the underlying assumptions of our experimental design is that valence is an accurate measurement of how much positivity is conveyed by a song. Based on exploratory data analysis as well as past work performed by Pimentel and Turnage (separately) of the research team, we see no reason to believe that valence is not an accurate measurement of positivity.

Using the Spotify API, we gathered song data and began to compile lists of songs with extremely high and low valence for the treatment and placebo playlists, respectively. The average valence values for our high and low valence playlists were .948 and .112, respectively. These extreme values were chosen to maximize power. The nature of the Spotify API does not allow us to search for songs by valence, so we were limited to using a guess-and-check method. These songs were picked specifically to have a good genre mix (i.e. a relatively uniform distribution of the most popular genres on Spotify) and to have a good mix of overall popularity.

Participants listened to all 3 of the songs back to back, immediately following completion of the pre-listening survey. Participants could control how long they stayed on each page (navigating to the next song required clicking the “next” button to get to the next page). We attempted to track how long each participant stayed on each page using Qualtrics, but ultimately found our measurements to be unreliable (see Non-compliance section for more detail).

After listening to the last song, participants were prompted to click the “next” button and then fill out the post-listening survey. The first question on this survey was the mood measurement question.

One question we commonly came across when presenting this methodology to others was, “Why did you choose low valence songs rather than “neutral” valence songs?” This is a fair question. Many people prefer to listen to sad or angry sounding music and find that songs of that nature boost their mood - more so than happy-sounding music! So are we mistaken in using low valence music as a placebo?

Our thought is that low valence is not inherently the opposite of high valence, but rather, that low valence represents an absence of valence. So while we use the terms “high-valence” and “happy-sounding” or “happy” interchangeably in this paper, we are in reality testing the effect on mood of exposure to music with high valence. To the extent this matches the definition of happiness is a subject for another study, but as mentioned previously, none of our work in this experiment or in prior analysis, nor in literature review, identified any reason why we should doubt this match. In the end, any significant results would only imply things about high-valence music, and we would leave it to further studies to conclusively prove a bridge from high-valence to happiness. We recognize that this may not be a universally accepted idea, and welcome thoughts surrounding the validity of this statement. We, however, feel that low valence music is an effective placebo for the phenomenon we are trying to study.

*Non-compliance*

We experienced extensive non-compliance that greatly hindered our power and the resulting value of our analysis. Initially, we did not expect much non-compliance so we only set up a control measuring whether someone visited every page. When we began to notice how quickly people were completing the survey, we attempted to set up controls that would record how long a participant was on a certain page by recording their first and last clicks on the page. We experienced bugs with this control that made us hesitant to rely on it for the purposes of measuring non-compliance, and we ultimately chose to not use the data gathered from this control.

In the end, we had two trustworthy metrics that we were able to use to estimate the rate of non-compliance:

1. “Progress” - a metrics we designed to track whether participants answered every question and visited every page
2. “Duration” - how long, in seconds, the participant was inside of the survey

We were immediately able to remove from the study participants for whom “Progress” did not equal 100%, that is, participants who did not visit every page. Estimating compliance using duration was a bit trickier.

As mentioned previously, participants were given 3 30-second long song snippets to listen to - for a total of 90 seconds. So we could use Duration to estimate whether participants actually got a “full dose” of treatment. To estimate how much of a dose participants received, we assumed each question on the survey took 5 seconds to answer. With 13 total questions on the survey, this came out to 65 seconds to answer all survey questions. So we were able to track “estimated compliers ” through the following formula:

**Formula 2**

Duration - (Number of questions \* time to answer each question) > 90

Admittedly, 5 seconds is a pretty arbitrary estimate. We found this to be reasonable given how long it took us to take the survey (we reasoned that participants took slightly longer than we did), but we concede that it is very possible that something closer to 10 seconds (or even higher) could be more reasonable. Without exact measurement, it is very hard to get a sense of a reasonable time per question due to the estimated high level of variability between participants. Further, we felt our usage of 5 seconds was appropriate given that we see insignificant effect sizes with this estimate - if we see insignificant treatment effects while assuming people got more treatment exposure than they actually may have, we find it unlikely that we would suddenly see significance with a longer duration assumption.

Upon running this check, we found substantial non-compliance. We considered several ways of moving forward with our analysis. We first considered partitioning our data into buckets based on exposure time, therefore creating buckets of levels of estimated exposure times. In the end, the complexities associated with analysis of the results of this method (including but not limited to dealing with and interpreting placebo partial compliance) caused us to re-evaluate our approach.

In the end, we made the assumption that those who we estimate received at least half of the total exposure time to treatment or placebo (i.e. listened to at least 45 seconds of music) would be treated as having fully complied. We treated everyone with estimated total exposure time < 45 seconds as a non-complier and removed their observations from our CACE analysis (this totaled 6 observations). This is not a perfect methodology, but we defend it as such: like our 5 second assumption, we believe this assumption risks underestimating rather than overestimating the CACE. To us, it is unlikely that those who listened to 45 seconds of music experienced as strong of an effect as those who listened to 90 seconds of music. Therefore, we feel our treatment of partial compliance is appropriate and conservative, although imperfect.

*Power*

As mentioned, we were unable to find any studies that had provided similar treatment and had a similar mechanism of measuring outcomes. Therefore, conducting a power analysis was tricky. Based on our literature review, we felt confident that our treatment would have a positive effect. As we measured outcomes using differences gathered from a pre and post-treatment 5 point Likert scale, we anticipated many of our outcome variables would be in the (-1,1) range. This implied a small treatment effect. We roughly estimated that we would see an effect size in the range (.25-.75) based on our research and literature review, with a standard deviation about double the size of the effect size - an estimate that we felt was conservative given the small amount of relevant and readily available research. With these measurements in mind (roughly μ1 = .5 , μ2 = 0, σ = 1) along with power = 1- β = .8, we estimated a necessary sample size of 126 (combined treatment and control). The team committed to recruiting 50 people each in order to account for non-compliance (although we ultimately failed to hit this number). Assuming no non-compliance, our target sample size of 150 would have given us power = 1- β = .86 given the above assumptions.

**Analysis**

*Data*

Our survey yielded a substantial amount of data. Ultimately, we used most of the data we captured from participants in our analysis. Below, we detail each variable that was used in our models (names in parenthesis represent that variable’s name in our dataset). A full list of variables is available in our dataset.

Change in mood (delta\_mood):

* Represents the difference in answers received for the mood questions (see Formula 1). This is the outcome variable in our analysis

Exposure to high-valence playlist (HVTotal):

* A binary variable representing whether the participant was assigned to treatment (1) or control (0)

Received full dose of treatment (Likely\_Full\_new):

* A binary variable representing whether the participant was estimated to have received a full dose of treatment (1) or not (0)

Received partial (45-90 second) dose of treatment (Partial\_1\_new):

* A binary variable representing whether the participant was estimated to have received a partial dose of treatment lasting between 45-90 seconds (1) or not (0)
* Note, observations representing a participant who received between 0-45 seconds of treatment are denoted by HVTotal = 1, Likely\_Full\_new = 0, Partial\_1\_new = 0

Hours of music listened to per week (Hrs\_music\_week):

* Average number of hours the participant spends per week listening to music

Hours of music listened to today (Hrs\_music\_today):

* Number of hours the participant has spent listening to music today

Favorite Song (Fave\_Song\_Played):

* Which of the three songs did the participant like the best?

Favorite Song Last (Fave\_Song\_Last):

* Binary variable representing whether the participant’s favorite song was the last one they heard

Least Favorite Song (Least\_Fave\_Song):

* Which of the three songs did the participant like the least?

Least Favorite Song Last (Least\_Fave\_Song\_Last):

* Binary variable representing whether the participant’s least favorite song was the last one they heard

How did the participant listen to the music (Mode\_of\_Listening):

* How did the participant listen to the music - on their computer, using a speaker, using their phone, using headphones, etc.

Similarity to the music the participant normally listens to (normally\_listen):

* Binary variable representing whether the participant would consider the music they listened to in treatment/control as similar to what they typically listen to

Did the participant hear their favorite genre of music (did\_hear\_favre\_genre?):

* Binary variable representing whether the participant heard a song from their favorite genre (this was not a question we asked in the survey, rather, we were able to deduce it from the participant’s answer to the favorite genre question as well as Spotify’s data on each song)

Average danceability (danceability\_mean):

* Continuous variable on (0,1) ranking the average danceability of the 3 songs the participants listened to. Danceability is a metric defined by Spotify as “how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.”

Average energy (energy\_mean):

* Continuous variable on (0,1) ranking the average energy of the 3 songs the participants listened to. Energy is a metric defined by Spotify as “a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.”

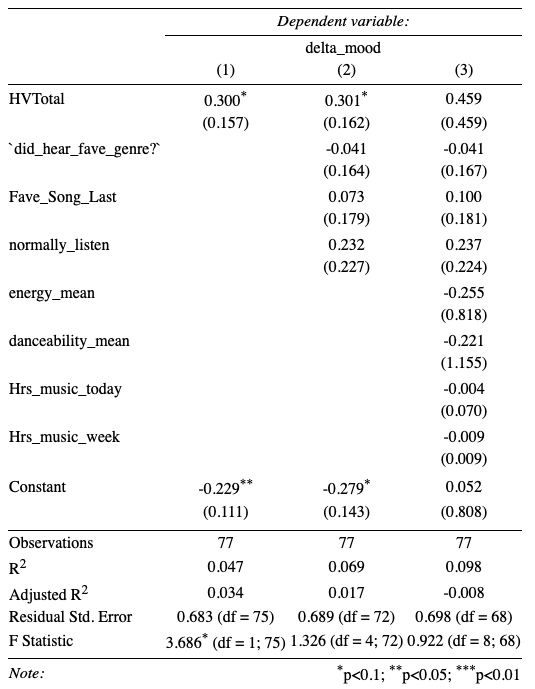
*Models*

Once we had gathered our data, we considered which variables made the most sense to use as covariates in our analysis. As a part of exploratory data analysis, we built numerous models regressing the outcome variable (the change in mood) on whether they received treatment (exposure to the high-valence playlist), and one of the variables detailed above as a covariate. This allowed us to narrow in on what seemed to be the most important covariates (although we should note that none of the covariates, nor the treatment variables, in any test, could be regarded as statistically significant at p < .05).

Given the fact that none of our tests yielded any significant results, the reader should not draw any conclusions about the relationship between happy music and mood (nor should they about any of the other covariates and their relationship with the treatment and outcome variables).

We hypothesize that with some changes, a similar experiment could yield significant results, and that some of the covariates we looked at are likely significant as well. These thoughts are only hypothetical - so we will not delve into the exact theory behind each relationship. However, for educational purposes, we will briefly discuss some of the more interesting variables and relationships in this section.

**Figure 2: 3 Models of Interest**



Results for each variable in each model are insignificant at the p < .05 level. Still, we feel it can be educational to walk through a brief discussion of the results, why we chose these covariates, and what these results could, in theory, mean.

From the base model, we find the following effects. Exposure to the placebo (low-valence) songs, on average, caused an effect of -.229 “points” on mood. This slight decrease seems roughly in line with what we believe we’d see given past research and our intuition.

The treatment effect itself was interesting. The treatment effect of exposure to the high-valence songs was .3 - meaning those in the treatment group had a very small, but positive, total effect (.071 since the intercept was -.229). While insignificant at p <. 05, it is significant at p < .10, which we believe warrants some sort of follow-up work. We note here that due to our placebo designs and treatment of non-compliance that this effect size estimate represents the CACE of the treatment (we discuss the ITT in Figure 3)

As for why we see a negative intercept: it could be that having been made to take our survey itself caused a decrease in mood. Taking a survey (even one as short as ours) is an inconvenience to many, and certainly not something that many people enjoy doing. It is not unreasonable to believe that participating in our experiment itself had an effect. Perhaps further research could leverage some sort of incentivization to offset this negative effect - although we understand that is a tricky thing to balance (since we also don’t want participation in our experiment to have a positive effect).

It also could be that people genuinely found the low-valence music to decrease their mood, while the high-valence music kind of kept their mood in check. Participants, on average, were already in a pretty good mood at the start of our experiment ( average moodpre = 3.75 for participants who met the requirements for full exposure. Maybe exposure to happy music, especially for such a short time, can only boost mood to a certain extent? This seems reasonable, especially when remembering that the total exposure time was only 90 seconds broken up over 3 songs. Listening to parts of different songs can be a little jarring, and even if they are very happy-sounding, they likely do not have as strong an effect as listening to a full song or several full songs. In our literature review, we did not identify any studies with such short exposure times spread out over multiple songs - most studies that had people actually listening to music had them listen to full songs (some even long classical pieces). We understand this is a limitation that could certainly have caused us to get a smaller treatment effect than normal, and is a potential limiter of our power.

We also note the greatly increased standard errors in model 3. Likely causes of this include the inclusion of covariates that do not give much additional explainability in a large model, although many seemed to offer decent explainability in EDA. We also believe multicollinearity between the audio features (valence, danceability, energy) could be causing this. All 3 have pretty large standard errors in model 3, and past work done by Pimentel has shown significant correlation between each of these 3 features in past years of songs featured on the Billboard Year End Top 100 songs. While work done in that study cannot be generalized to explain audio feature relationships here, it is not entirely surprising that we see these effects in model 3.

The covariates in model 2 and 3 were carefully picked after examining response distribution and running several models in EDA. These models did not yield any significant effects at the p < .05 level. As it turns out, their inclusion did not meaningfully increase model performance, and seem to have even decreased model performance. We can see this in both the ballooning standard errors of our treatment effects (really only in Model 3), decreasing adjusted R2, and decreasing F statistics. Although we did not see any significant effects, we briefly outline our theories for why these covariates could affect the change in one’s mood below:

We hypothesized that hearing a song from one’s favorite genre could cause an increase in mood. If you really like rap music, then, independent of whether you are listening to a happy-sounding or sad-sounding rap song, it is intuitive that your mood might increase more than listening to music from a genre you are impartial to. We did not find any evidence to back up this claim, and we actually saw a slightly negative point estimate for having heard one’s favorite genre (although with relatively large standard error)

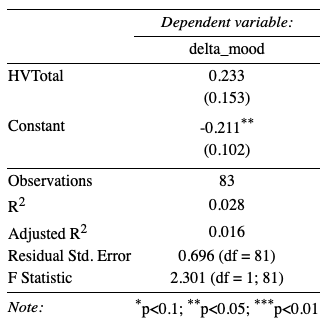
We hypothesized that if your favorite song out of the three was the song that you heard last, then maybe you would experience an increased change in mood. The theory behind this is that perhaps just the last song or last x seconds/minutes of a song is what impacts mood the most. If we had seen a significant positive effect here, similar studies could be done to examine just how much exposure time is needed to see an effect on mood. We did not find any evidence to back up this claim.

Similar to our genre hypothesis, we also hypothesized that if you classified the music you listened to in treatment/control as similar to what you normally listen to, then you may have an increased benefit to mood. We did not see any evidence to back up this claim, although we did observe a positive effect size.

We also hypothesized that maybe it is not only valence that causes a change in mood, but rather how “danceable” a song was or how what level of energy was conveyed in the song. It is intuitive to think that songs that make you want to dance or give you energy could also boost your mood. We ran models that regressed delta\_mood on both danceability and energy alone in EDA and saw positive effect sizes with p-values similar to our simple, valence-only model (Model 1 of Figure 2 above. Although insignificant (average danceability had an estimate of .752, standard error of .452, p = .100, average energy had an estimate of .697, standard error .512, p = .178) these would be interesting areas for future study.

Finally, we hypothesized that the amount of music a person listens to per week, as well as the amount of music that person had listened to that day before taking the survey, would both impact the change in mood. It could be that those who listen to a lot of music every week experience increased benefits to mood (which could explain why they listen to so much music). We also thought it could be the case that those who had listened to a lot of music that day would receive lower marginal benefits to listening to further music. We did not find any evidence to back up either of these claims.

Ultimately, we believe that we picked good covariates (perhaps besides mean danceability and mean energy) and would likely pick the same or similar covariates in a future study. We think that investigating methods to increase power or limit non-compliance (as we discuss elsewhere in this paper) would be key to seeing potential true effects of the treatment variables and covariates.

**Figure 3: Measuring the ITT**

In measuring the ITT, we used the results of all observations for which delta\_mood was defined. This ended up being 81 observations. Although results are insignificant, similar to our analysis in Figure 2, we see a negative effect of receiving placebo, and a similar sized positive effect of treatment. In general, there is a narrower spread between treatment and control. This does not really give us any more information than our analysis of the CACE does - except to say that it seems that non-compliance with treatment was associated with lower increases in mood. This makes sense, given non-compliers are not receiving a treatment that would, in theory, boost their mood.

**Discussion**

As we were unable to obtain any statistically significant results, we feel it is inappropriate to hypothesize on any causal mechanisms, covariate relationships, or specifics of the treatment effect in greater detail than we have already done. Any further discussion would verge on hypothesizing according to intuition and prior research, rather than any research completed here. Instead, we wish to discuss some shortcomings with our process and work that we believe can inform future research.

*Power*

Obtaining sufficient statistical power was a concern from very early on in this project. Much of the research we reviewed prior to starting this experiment was very closely adjacent to what we intended to do, but the experiments and measurement methodology were different enough that our power calculation was just a rough estimate. We discuss the specifics of our power calculation in the Power section of the Experimental Details portion of the paper. Ultimately, non-compliance was a much more significant problem than we originally anticipated, which heavily decreased our power(see Recruitment / Non-Compliance section below).

We also believe it is possible that we could have gotten more power by using a larger scale to measure mood - perhaps a 10 point rather than 5 point Likert scale. While our initial research did not focus on this, we believe that it is possible that framing mood as a 5 point scale actually caused us to capture less information than a 10 point scale would have gotten us. We think it is possible that people think of mood in a broader range than the 5 point scale allows for. For example, maybe people don’t feel quite at a 3 or a 4, but actually feel like a 3.5. Our study had no way of capturing this data. Increasing the range of possible responses would certainly raise standard errors, but we think it is reasonably possible that this increase in standard errors due to increased variance in responses would actually have less of an effect on power than that of increasing the range of potential responses.

*Recruitment / Non-Compliance*

As we did not have the funds nor the time available to engage in a sophisticated random sampling process, we relied heavily on convenience sampling. As we note elsewhere, we did also post links to our survey on social media sites, but based on the low amount of engagement with these posts, we do not think that we had many participants that came from those links. If we had found significant results, it would be difficult, if even possible, to reason about how these effects might generalize to a broader population given our sampling methods. As we did not obtain results that make that discussion informative beyond a simple discussion of sampling theory, we will refrain from discussing these generalizations here.

In order to accurately measure the phenomenon we attempted to measure in this experiment, doing so in a more formal environment, with well-thought out preventative measures, is probably necessary. Doing so, of course, probably requires some level of funding, some level of support from a group that has experience collecting participants and running studies, and a significant amount of time.

Partly because of these and other limitations, our effect size estimates were calculated using an estimate (see Figure 2) that is almost surely incorrect. Even if we were able to accurately measure how long each participant listened to each song, we have no way of knowing, for example, whether they are actively listening, or if they have just left the song playing on their laptop while they are in another room. Conducting any sort of experiment where the researchers are unable to observe the participants leaves open the possibility of non-compliance - in ways and amounts that are often difficult to predict.

Further, when dealing with partial levels of non-compliance, making a non-arbitrary partition of the data into meaningful buckets is difficult. We chose to assume an exposure time of > 45 seconds was equivalent to being fully treated. We recognize this partitioning is arbitrary and that better partitions, along with better methods of analyzing them, likely exist.

We discuss in further detail the non-compliance issues we observed and our chosen methods for dealing with these issues in the Non-Compliance section of the Experimental Details portion of our paper.

*Conclusion*

It is unfortunate that we were unable to identify any significant effects, especially because the team does believe that happy-sounding music should have a significant effect on mood. Still, we feel that running this experiment has been a fantastic learning experience and that we walk away from this experience with knowledge on how to run a better experiment in the future. We hope this discussion has been informative and welcome any and all questions the reader may have, or suggestions for further improvements for future research.

1. Impact of Music on Mood: Empirical Investigation (Ahmad, Rana 2015) [↑](#footnote-ref-0)
2. Effects of music interventions on stress-related outcomes: a systematic review and two meta-analyses (de Witte, et. al 2018) [↑](#footnote-ref-1)
3. Trying to be happier really can work: Two experimental studies (Ferguson, Sheldon 2013) [↑](#footnote-ref-2)
4. The Rewarding Aspects of Music Listening Are Related to Degree of Emotional Arousal (Salimpoor, et. al 2009) [↑](#footnote-ref-3)
5. Music Alters Visual Perception (Jolij, Meurs 2011) [↑](#footnote-ref-4)